Applications of Risk Pooling for the Optimization of Spare Parts with Stochastic Demand Within Large Scale Networks

By

Nigel Goh Min Feng

B.Eng. Mechanical Engineering,
The University of Western Australia, 2013

Submitted to the MIT Sloan School of Management and the MIT Department of Mechanical Engineering in partial fulfillment of the requirements for the degrees of

Master of Science in Mechanical Engineering
and
Master of Business Administration

in conjunction with the Leaders for Global Operations Program at the

Massachusetts Institute of Technology

May 2020

Author ....
MIT Sloan School of Management, Department of Mechanical Engineering
May 08, 2020

Certified by ....
Nikos Trichakis, Thesis Supervisor
Associate Professor, Operations Management

Certified by ....
Kamal Youcef-Toumi. Thesis Supervisor
Professor, Mechanical Engineering

Approved by...
Maura Herson
Assistant Dean, MBA Program, MIT Sloan School of Management

Approved by...
Nicolas Hadjiconstantinou
Chair, Mechanical Engineering Committee on Graduate Students

This page intentionally left blank.

Applications of Risk Pooling for the Optimization of Spare Parts with Stochastic Demand Within Large Scale Networks

By

Nigel Goh Min Feng

Submitted to MIT Sloan School of Management and the MIT Department of Mechanical Engineering on May 08, 2020 in partial fulfillment of the requirements for the degrees of Master of Business Administration and Master of Science in Mechanical Engineering

Abstract

Amazon is able to deliver millions of packages to customers every day through its Fulfillment Center (FC) network that is powered by miles of material handling equipment (MHE) such as conveyor belts. Unfortunately, this reliance on MHE means that failures could cripple an entire FC. The exceptionally high stock-out cost associated with equipment failure means spare parts must always available when required.

This is made difficult as Amazon does not hold any central repository of inventory at present – all inventory is held at a site-level. Unfortunately, FCs have to stock more inventory than required due to unpredictable failures, long lead times from suppliers, and no standard work processes for site-to-site transfers. However, if Amazon is able to pool its spares across multiple FCs, it has an opportunity to reduce the spares kept across the entire FC network, position itself to better respond to catastrophic failures, and consolidate interfaces with suppliers.

The goal of this thesis is to identify the inventory model and network design that would maximize parts availability while minimizing cost. Additionally, an implementation roadmap will be developed to outline how such a system (e.g. hub locations, logistic channels etc.) can be developed. This thesis concludes by proposing potential extensions of the work conducted in this thesis to improve the practicality and financial impact of the proposed network and inventory model.

Thesis Supervisor: Kamal Youcef-Toumi
Title: Professor of Mechanical Engineering, MIT

Thesis Supervisor: Nikos Trichakis
Title: Zenon Zannetos (1955) Career Development Professor, MIT Sloan School of Management

The author wishes to acknowledge the

Leaders for Global Operations Program for its support of this work.

Acknowledgements

I would like to first thank my MIT academic advisors, Professors Nikos Trichakis and Kamal Youcef-Toumi, who both took significant amounts of time out of their busy schedule to provide guidance and direction throughout my internship. This thesis would not be anything close to what it is if not for their help, and I am tremendously privileged to have had the opportunity to learn from them.

I would also like to thank Amazon for the opportunity to work on what is an essential part of their operations, and for the support that they have provided through the entire process. I want to also specifically thank John Fogerty, who was both manager and friend, for his trust and constant feedback. Additionally, I’d like to thank Brent Yoder who championed my project from start to finish, and was instrumental in ensuring that the project stayed on track and never ran into problems.

Next, I can’t possibly write acknowledgements without writing about my LGO classmates – an amazing group of people who keep me grounded and humble every day. I am extremely thankful for their friendship and for making these two years at MIT a remarkable experience that I will always remember.

Finally, I would like to thank my family. I’m only where I am today because of their love, encouragement and support. Thank you for helping me (try to) be the best person that I can be. I am, and will be, forever grateful to them for everything they have done for me.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Project Background

Amazon is able to fulfill its two-day delivery promise and deliver millions of packages to its customers every day through its Fulfillment Center (FC) network, where FCs hold inventory that is available for sale on Amazon. When a customer places an order online, the order is assigned to an FC that has the item in stock, and the item is then packed and shipped to the customer. The FC network is powered by miles of material handling equipment (MHE), such as conveyor belts and rollers, that help to move the inventory across the FC.

In order to ensure that customers are able to receive their packages on time, FCs are required to hold spare parts for all MHE. This ensures that any breakdowns can be immediately repaired and will not disrupt any flow of inventory within, and out of, the FCs.

1.2. Problem Statement

Amazon’s ability to deliver on its two-day promise to customers is highly reliant on the continuous operation of Material Handling Equipment (MHE) within the various Fulfilment Centers (FCs). This dependence on MHE means that equipment stock-outs could have significant consequences for operations. As such, Reliability and Maintenance Engineering (RME) must ensure MHE spares are always available when required. However, holding an excessive amount of MHE parts is neither feasible nor desirable due to space and cash flow constraints.

Inventory levels are currently controlled through a Min/Max inventory policy. However, the Min/Max levels are set based on vendor recommendation and site admin experience, not usage data of the parts. Variability in demand for parts and long lead times (typically four weeks or more) make it difficult to accurately determine optimal stocking volumes for MHE spares. Comparisons against a data-based inventory model show that only 30.2% of current parts are stocked at an optimal level. This means that many NAFC sites are either holding excess inventory or not holding enough inventory (increasing the risk of a stock-out).

The table below show how 102 NAFC sites (with >500 parts in stock) perform in terms of having a certain proportion of their SKUs within percentage bands of "optimal" stocking values based on best practice inventory policies. This data is not included as a precursor for further analysis on current stocking policies, but rather seeks to simply shed light on the potential of this program.

Correct % of SKUs	# of Sites
< 15%	7
15% – 20%	9
20% - 25%	14
25% - 30%	21
30% - 40%	31
40% - 50%	19
> 50%	1

Table 1.1: Current Stocking of SKUs

This is further complicated as Amazon does not have a central repository for inventory, so parts are ordered directly from suppliers. Each FC manages its own budget and inventory through an Enterprise Asset Management (EAM) software and a dedicated site-based EAM administrator who is responsible for the spares cage in the FC. Since FCs are managed at the site-level, and not as a network, it is difficult to identify duplicate parts between sites, consolidate suppliers (and leverage Amazon's buying power), and better respond to emergencies and stock-outs by sharing parts between sites in the network.

1.3. Research Hypothesis and Methodology

This presents two opportunities to improve FC operations across the entire network. First, Amazon has significant amounts of data regarding spare parts usage that could be used to better inform stocking levels of spare parts. Second, risk-pooling between sites can be implemented to share and manage parts at the network level.

Risk-pooling creates a multi-echelon supply chain by introducing centralized hubs within the FC network that FCs can source parts from. This significantly lowers part lead times at an FC-level and also pools the demand for spares across multiple FCs – both of which help to reduce inventory levels. This “nodal warehousing” strategy presents an opportunity for Amazon to not only reduce the spares kept across the FC network, but to also position itself to better respond to catastrophic failures whilst consolidating interfaces both internally (e.g. inventory planners) and externally (e.g. vendors).

As such, there were two key goals for this project. First, to identify the optimal risk-pooling design for Amazon’s FCs and to quantify any potential savings that would result from its implementation. Second, to use data to determine the optimal part stocking levels that would maximize part availability while minimizing cost.

1.4. Project Scope and Limitations

Due to the scale of Amazon’s operations, in order to ensure that the project could be done within the available timeframe, only Amazon’s North American operations were considered. Although the scope of work from this project could be extended to other geographies, there are other factors that will have to be considered when doing so (e.g. cross-border complexities between countries).

A core part of this project involves the use of data regarding spare part usage and ordering. One limitation of this data is that it is extracted from a system that involves manual input from users, and there are observations where those manual inputs are incorrect. However, due to the number of data entries within the system, it is not feasible for the data to be manually filtered. As such, systems will be put in place to identify and filter out incorrect data entries where possible, but all other data entries will be taken as accurate.

This study also assumes that inventory levels cannot be run down to zero (i.e. where FCs use a pull system when spares are required) as that would necessitate downtime whenever a spare is required. As the priority for FCs is to ensure that there is no downtime, this is considered to be not acceptable.

Finally, although Amazon made all its data available as part of this thesis, for data privacy reasons, all data that is specific to Amazon was presented to Amazon at the conclusion of the project and will not be reproduced within this thesis. Instead, this paper will focus on the process and general learnings that are applicable to other scenarios.

2. Literature Review

Spares management and inventory modelling are well-established areas of operations research. This chapter aims to provide an overview of the prior work already done in this space and to establish the context required to understand the other chapters of this paper. The literature review then also looks into potential ways in which risk-pooling can be applied to supply chains, and provides a glimpse into the facility location problem – the problem of placing warehouses in an optimal location.

2.1. Inventory Theory

One key challenge for any operation involving variable demand (e.g. retailers, warehouses etc.) is determining the quantity of inventory to hold. If the operation does not hold sufficient inventory, they risk running out of stock which could result in lost sales or downtime. On the other hand, excess inventory ties up cash flow and space, both of which could be used to improve the operation in other ways, and typically has associated holding costs.

Another inventory management challenge lies in restocking of inventory. Inventory typically has to be ordered in advance of when they are required due to lead times from suppliers (which typically adds another dimension of variability). This is further complicated as inventory will be further consumed while waiting for new inventory to be delivered from suppliers. As such, inventory managers have to determine how much inventory to order, and when to order that inventory before they have sufficient data to do so. Due to the highly visible consequences of stockouts, there is a temptation for warehouses/retailers to hold more inventory than required.

In response to these challenges, mathematical inventory models have been developed to improve inventory policies which provide guidance on timing and quantities of inventory replenishment. Although there are many different models, there are several distinctions that are of particular interest to this paper:

1) **Deterministic vs Stochastic Demand:** Demand profiles will vary depending on the type of function the inventory is used for. If future demand is well-known and can be accurately forecasted, a deterministic inventory model is used. However, if future demand is a variable rather than a known constant, then a stochastic inventory model has to be used (Jensen & Bard, 2002).

   Deterministic demand profiles consume inventory continuously at a known and constant rate, whereas stochastic demand profiles have uncertainty in the demand. In both cases, inventory is replenished when needed by ordering a certain replenishment quantity. (b)

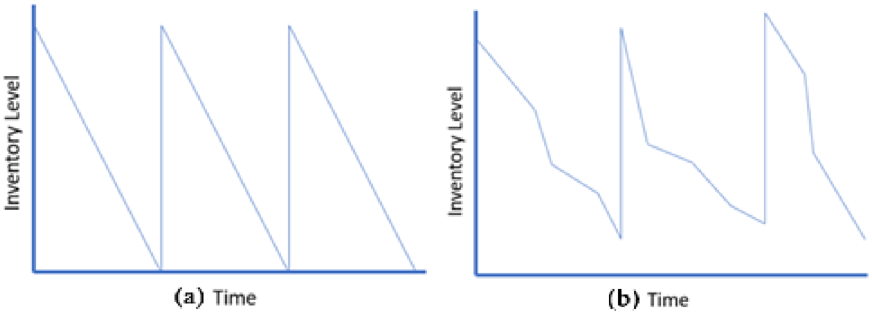


Figure 2.1: (a) Deterministic vs (b) Stochastic Demand

2) **Single vs Multi Period:** An inventory model has to determine the correct amount of inventory to stock in order to optimize costs and/or profits. However, the model may be used for a single period (i.e. once the period is over, the parts are no longer required), or for multiple periods. A single period considers inventory as independent of future periods, whereas multi-period models carry over inventory from period to period, which complicates the inventory policy (Zhang et al., 2009).

3) **Periodic vs Continuous Review:** For multi-period inventory models, any consumed inventory will typically have to be replenished at some point in time. A periodic review model checks the inventory level at specific intervals, and replenishment orders are only made during those time windows. In a continuous review model, inventory is continuously monitored, and an order is

placed as soon as the inventory level falls below a certain threshold known as the reorder point (Zappone, 2006).

For the purposes of this paper, Amazon faces a stochastic demand over multiple periods, but is able to implement a continuous review process.

2.2. Inventory Models

2.2.1. Newsvendor and Economic Order Quantity

Any operation with fluctuating demands will typically face issues as the demand uncertainty could result in overstocking and stockouts. Numerous inventory models have been developed over the years to deal with this and the newsvendor model has been a mainstay of inventory theory since it was developed in 1951 by Morse and Kimball.

The newsvendor model applies to scenarios where an operation faces stochastic demand and has to determine its required quantity of inventory before knowing the required demand (Morse & Kimball, 1951). The overall objective of the newsvendor model is to identify the inventory quantity that best balances the overage and underage costs through statistical information on the demand (e.g. the mean and standard deviation of the demand). Using the newsvendor model, it is possible to determine a level of carrying inventory that is required to meet the expected demand for a given period of time.

The newsvendor model typically assumes a normal distribution due to its ease of application. However, due to a high probability of negative demands (which is not theoretically possible) when the mean is low and variability is high, the normal distribution may not be a reasonable approximation of the demand all the time. As such, in the event of a high coefficient of variation of demand, typically defined as >0.5, one should consider a distribution other than a normal distribution (Halkos & Kevork, 2012; Silver et al., 1998).

Although the newsvendor model is able to provide operations with a sense of how much inventory is required for a single period, it does not help to determine how much inventory to reorder to make up for consumed parts. The Economic Order Quantity (EOQ) was first developed by Ford Whitman Harris in 1913 and can be applied to inventory items that are replenished in batches. The EOQ model considers two cost buckets, holding costs and ordering costs, and aims to minimize the combined cost of both (Goh, 1994).

2.2.2. (R,Q) Inventory Model

Although the newsvendor model and EOQ provide a baseline upon which an inventory policy can be built, there are more specialized inventory models that are more applicable to Amazon. The (R,Q) inventory model typically involves a continuous review period over multiple periods, and is better suited for Amazon's purposes.

The (R,Q) policy assigns a fixed replenishment point and a fixed replenishment quantity for each part. The (R,Q) model has two parameters: whenever the inventory on hand falls below a certain Reorder Point (R), it will trigger a new order for a specific Reorder Quantity (Q). The reorder point is set such that there is sufficient inventory on hand to meet all reasonably expected demands while waiting for the replenishment quantity from the supplier. Reorder quantities are set to minimize the total cost of ordering and holding inventory through the EOQ (Cachon & Terwiesch, 2013; Capar & Eksioglu, 2009).

Reorder Level (R)

The Reorder Level is the inventory quantity at which new inventory is ordered. Intuitively, when a new order is placed, the system must ensure that there are sufficient parts on hand to cover demand until the new parts arrive. Accordingly, this Reorder Level for any part corresponds to the sum of the Cycle Stock and the Safety Stock. These terms are described further below:

1) **Cycle Stock** is the inventory that is expected to be used for any given period of time. This time interval is typically the lead time for the part (i.e. How many parts will be used while waiting for new parts to arrive).

$$Cycle\ Stock\ [Parts] = Average\ Demand\ per\ Time\ (\mu)\ \left[\frac{Parts}{Time}\right] * Lead\ Time\ [Time]$$

2) **Safety Stock** refers to the additional inventory that is kept on hand to account for weeks that have larger than expected demand requirements (as it is impossible to predict the exact amount used every week).

$$Safety\ Stock\ [Parts] = Standard\ Deviation\ (\sigma)\left[\frac{Parts}{Time}\right] * \sqrt{Lead\ Time}\ [Time] * Safety\ Factor$$

The Safety Factor is the z-score corresponding to a part’s desired service level on the normal distribution (i.e. a Safety Factor of 3 will ensure sufficient inventory to cover inventory requirements 99.7% of the time -- a 99.7% service level).

Reorder Quantity

As more parts are ordered, the “cost per part” decreases due to fixed costs being spread across more parts. However, this is countered by the increased cost of holding more inventory. The Reorder Quantity (Q) aims to minimize the total cost of placing an order by optimizing for both ordering costs and holding costs.

$$Q\ [Parts] = \sqrt{\frac{2 * Demand\ \left[\frac{Parts}{Time}\right] * \ Fixed\ Ordering\ Cost\ [\$]}{Holding\ Cost\ [\frac{\$}{Time * Part}]}}$$

Whilst most parameters are fixed, the holding cost may be set based on numerous factors. Berling presents a general model based on microeconomics through which a holding cost can be determined (Berling, 2008).

2.2.3. (S-1, S) Inventory Model

An alternate model to the (R,Q) inventory model is the base stock model, also known as the (S-1, S) model. The (S-1, S) model works like the (R,Q) model in most ways. The main difference lies in how parts are replenished. The (S-1, S) model sets the required inventory level at (S), and whenever parts are consumed, it immediately replenishes its inventory back to the required inventory level of "S" (Cachon & Terwiesch, 2013). This is also known as an "order up to" inventory model, as it will replace any consumed inventory up to the required threshold.

The (S-1, S) model can also be thought of as an (R,Q) model with a reorder level of (S-1) and a reorder quantity of 1.

2.3. Inventory Pooling

Inventory pooling is a well established branch of operations research that was first introduced by Eppen in 1979 (Eppen, 1979). The concept of inventory pooling has been further explored since 1979, and it has been shown that pooling works across different demand distribution types (Federgruen & Zipkin, 1984). At a high level, inventory pooling combines various demand streams together in order to minimize the effects of demand uncertainty as the pooled demand will help to balance high and low demand variations (Bimpikis & Markakis, n.d.).

Graves and DeBodt show that it is possible to extend the traditional (R,Q) inventory model to a multi-echelon supply chain with reasonable accuracy (DeBodt & Graves, 1983). Additionally, Axsäter shows that, for multi-echelon supply chains with low demand, it is suitable to apply continuous review policies as opposed to needing to use a periodic review policy for high demand items (Axsäter, 1993).

Multi-echelon systems do **not** reduce the average demand of its constituent FCs. If the system needed 10 parts per week, it will still use 10 parts per week after nodal warehousing is introduced. As such, weekly consumption of Cycle Stock, which is defined through demand levels, are not affected by pooling.

The benefit of pooling comes from reductions in demand variability which reduces Safety Stock levels. Safety stocks at an FC-level drop due to a lead-time reduction. At a systems level, safety stock reductions occur by combining variations. Assume that there are three sites, each with a part with a standard deviation of 5. Using the formula above, the “combined” standard deviation is 8.66 (as opposed to the total standard deviation of 15 if they were kept separate). This almost halves the amount of safety stock that need to be held within the network to cover for demand fluctuations.

3. Methodology

This chapter details the research expectations of this project, the steps required for the data analysis and inventory model development, the risk-pooling methodologies, and the various considerations that went into improving and selecting the final model.

3.1. Research Hypothesis

As failures can occur at any time, there is much uncertainty surrounding the rate of spare part consumption. This is one of the major factors that complicates spare parts stocking levels at Amazon FCs. Given Amazon's desire for 100% uptime on MHE, FCs must ensure they have sufficient spare parts on-hand at any given time. Unsurprisingly, this could result in many FCs holding more parts than required. This study expects that inventory levels across the FC network can be optimized in two ways.

First, historical data on parts consumption can be used to advise stocking quantities. The EAM system has access to part data (e.g. price, lead times etc.) and demand profiles for each part over the last five years. The demand profiles can be used to determine an expected mean and standard deviation for the consumption of each part over a given period of time. These parameters can be used alongside the available part data to determine an optimal stocking quantity for each part through an inventory model, such as the (R,Q) model described in Chapter 2.

The second way is through risk-pooling across sites through a multi-echelon supply chain to minimize demand variations across the network. At present, all FCs individually manage their inventories with minimal sharing between sites. This exposes every FC to high variability in week-to-week part consumption. These demand fluctuations result in sites needing to stock enough parts to deal with weeks of high part consumption, even when those weeks of high demand do not happen often. A central warehouse for risk-pooling will normalize weekly demand across the entire network, and should result in lower carrying volumes of spare parts.

The central warehouse should stock all the required spared, and FCs can order parts directly from the central hub. This will dramatically shorten lead times for FCs to a matter of days. As shown by the equation for the reorder point, a shorter lead time will lower the quantity of parts that each site has to hold.

$$Cycle\ Stock\ [Parts] = Average\ Demand\ per\ Time\ (\mu)\ \left[\frac{Parts}{Time}\right] * Lead\ Time\ [Time]$$

$$Safety\ Stock\ [Parts] = Standard\ Deviation\ (\sigma) \left[\frac{Parts}{\sqrt{Time}}\right] *\ \sqrt{Lead\ Time}\ [Time] * Safety\ Factor$$

The network as a whole does not benefit from the shorter lead time (as the central hub still has to order from suppliers with the same four to six week lead time), but benefits from pooled demand as combined variance/standard deviation is lower than the sum of its parts as shown below:

$$\sigma_{pooled} = \sqrt{\sum_{i=1}^{n}(\sigma_1^2 + \sigma_2^2 \ldots + \sigma_n^2)}$$

Intuitively, this can be understood as a higher week of consumption in one site will, when spread over hundreds of sites, be balanced out by a week of lower consumption in a different site. This will lower the required safety stock that needs to be held across the network. Note that the cycle stock (i.e. the amount that is expected to be consumed every week) is not affected by this change. There is also a possibility that centralized stocking and ordering will provide Amazon will more market power to negotiate shorter lead times with suppliers, but we assume its effect to be zero for the purposes of this study.

3.2. Data Analysis

The EAM data forms the foundation upon which the rest of the project is built. Although a large amount of information is available on Amazon's servers, not all the information is relevant. Accordingly, any necessary data must first be identified, before then extracting the data from Amazon's servers through SQL queries. The accuracy/completeness of these queries must then be validated to ensure that

As much of the data on EAM was filled in manually, there is a non-zero possibility of inaccuracies in the data. The data thus has to be filtered to determine what can be used, and what has to be excluded from the study. Due to the size of the data sets used, it was not possible to validate the data at an individual level. Issues were identified by trending and aggregating data, which highlighted areas that needed further investigation. For example, parts that were returned late resulted in choppy data that skewed weekly consumption numbers.

These data trends provided valuable insights into not only data inaccuracies, but also key areas for improvement. For example, there were specific part types that accounted for a large proportion of weekly demand, and warranted special attention and focus. However, for data privacy reasons, the results from this section are not explored as part of this thesis as they do not provide any general learnings. Instead, they formed part of the internal recommendations that were made available Amazon through this thesis.

Finally, in order to maximize the applications for this study, data that was missing from EAM was estimated where possible. Only parameters that could be ascertained with a high degree of confidence (e.g. FCs may not have a price assigned to a part, but the part's price is known across the network) was included in the study.

3.3. Model Development

Once all of the data is verified and made available, it is possible to then fit the parameters for every part into an inventory model. The (R,Q) model was chosen to establish a specific inventory policy for every part in FCs due to its ease of use and continuous review period. First, the study considers how an application of the (R,Q) inventory model will affect inventory levels without the use of risk-pooling. This will enable effective evaluation of the impacts of applying an inventory policy by comparing the new inventory numbers to the current state. Additionally, this will also establish a baseline to evaluate the impacts of risk-pooling and whether it is worth pursuing. For example, if risk-pooling only reduces inventory levels by 2%, it may not be worth the capital and effort required to set up the multi-echelon model.

Once the baseline (R,Q) model is established, it is possible to explore other inventory models and perform sensitivity analyses to identify key parameters that affect the overall inventory policy and where the biggest rooms for improvements are. Again, for data privacy reasons, the results from this section are not included in this paper as they do not provide any general learnings.

The final step involves the inclusion of risk-pooling into the overall network through a multi-echelon system (also called nodal warehousing system in this thesis) that lowers inventory levels in the system by pooling variability. The key outcomes desired from this phase of the project was a determination of the number of nodal warehouses to be used, as well as the locations of those warehouses.

A singular nodal warehouse will "pool" the most parts together, and should result in a lower inventory level. However, these gains could be undermined by longer lead times to FCs (a singular warehouse would increase the net distance between hubs and FCs) and/or higher shipping costs. For the purposes of this study, it is assumed that the shipping cost from suppliers to the hub is equivalent to existing shipping costs. As such, any incremental shipping cost will be the result of the extra shipping leg between the hub(s) and the FCs.

First, potential hub locations have to be identified, and the distance between those hub locations and FCs have to be mapped (to accurately predict shipping costs and lead times). Next, the demand parameters for each FC/hub need to be determined. In multi-echelon inventory, each FC is considered a separate entity. Fortunately, the (R,Q) model also applies to multi-echelon systems. The only difference is that the hub's inventory is not simply the physical inventory on-site, but also includes all inventory downstream of it. This formula is applied to each part to ascertain the overall network inventory level for every part:

$$Hub\ Inventory\ Level = \ Parts\ at\ Hub + \sum Parts\ at\ FCs + \sum Parts\ in\ Transit\ to\ FCs$$

For example, if a hub has one motor on its shelves, and it supports two sites, each holding one motor, the inventory level of the hub is three motors. It is this combined inventory that triggers the Reorder Point.

The lead time for the hub, for the purposes of (R,Q) model implementation, is given by the maximum amount of time it takes to get from a supplier to an FC (through the hub):

$$Hub\ Lead\ Time = Time\ from\ Supplier\ to\ Hub + \max(Lead\ Time_{hub-to-FC})$$

Essentially, the hub and its FCs are treated as one "giant site" with the combined demands of all its constituent FCs. For example, assume a hub supports two FCs, each requiring one part a week. Assume also that it takes four weeks for a part to get from supplier to the hub, and it takes one week to go from the Hub to the FCs. This is the same as one giant site that uses two parts a week, with a lead time of five weeks.

Demand profiles are treated in a similar way, except the hub has no demand of its own. Pooled demand parameters are calculated as follows:

$$Average\ System\ Demand = \sum_{n=1}^{All\ FCs} Average\ Demand\ of\ Part\ at\ FC_n$$

$$System\ Standard\ Deviation = \sqrt{\sum_{n=1}^{All\ FCs} (Standard\ Deviation\ of\ Part\ at\ FC_n)^2}$$

These parameters allow for the (R,Q) model to be applied to every possible hub configuration (both location and quantity) to identify the best configuration. As nodal warehousing will increase shipping costs (due to additional shipping between the central warehouse and the FCs) but reduce inventory volumes, the optimal solution would be defined by the most significant reduction in overall cost.

3.4. Summary

Inventory models can provide guidance on inventory stocking policies at an individual FC level. In the particular case of Amazon, the (R,Q) model was chosen due to its continuous review period and easy applicability. After the (R,Q) model is applied to each FC, the entire network can be further optimized through risk-pooling, which can further reduce inventory levels by reducing overall demand variability across different sites.

4. Data Breakdown and Analysis

As shown in Chapter 3, a large amount of data is required to apply the required models. This chapter details how that data was managed as part of this thesis. It explores the available data from EAM, how missing parameters were estimated based on available information, and how data accuracy was improved through systemic elimination of outliers. It then looks at how the final dataset can be aggregated and used to draw further conclusions regarding the overall state of the supply chain.

4.1. Current State

The bulk of the data used for this project was extracted from Amazon's Enterprise Asset Management (EAM) software which captures most of the information regarding Amazon's spare parts. However, the data on EAM is not populated automatically, and instead relies on EAM admins to manually input most parameters (e.g. price of part, quantity purchased/consumed etc.) In the past, EAM served as a platform for tracking data, but this data was not always utilized. As such, there was no incentive for EAM admins to ensure 100% data accuracy on EAM. This has resulted in some parts not having all the data required for further analysis (e.g. price, lead time, usage data etc.) and other parts having inaccuracies within the data (e.g. prices may be missing a zero).

Data inaccuracies were a larger concern that missing data as any inaccuracy could result in significant deviations from the ideal inventory policy for any given part. For example, if the price of a part were $10, but was input into EAM as $100 by accident, that could result in dramatically lower stocking quantities due to the higher price of the part. However, as there were millions of data points, it was not feasible to manually inspect each data point to identify problematic data. Instead, data was aggregated and trended to identify outliers, which could then be manually inspected and removed if necessary.

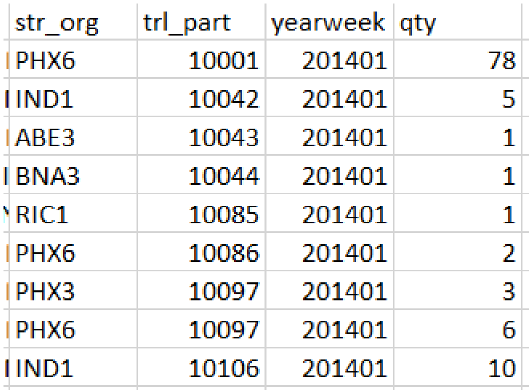
The remainder of this chapter details the data that was available, and how the above steps were achieved.

4.1.1. Available Data

On top of having part-specific data (e.g. price), EAM also records every instance where a spare part is used within any FC. Although usage data for part consumption was available on a daily level, it was decided that data on such a granular level was unwieldly and unnecessarily detailed for the analysis that this project had to perform. Instead, data was compressed into weekly consumption levels for better interpretability and to minimize the impact of potential outliers by lumping data together. For example, if a site (e.g. BFI4) were to use one part on Monday, and two parts on Thursday of a given week, the system would record BFI4 as having used three of those parts in that week.

The table below shows the most relevant information that was extracted from EAM through SQL queries and how they were important to the overall model development process.

<table>
<tr><th>Parameter</th><th>Description</th></tr>
<tr><td>Part Number</td><td>Every part has a unique part number that identifies the part. This number is consistent across sites and allows for aggregation across sites.</td></tr>
<tr><td>Site</td><td>Parts are used across sites and it is important to identify where the data is coming from (sites will have different prices/consumptions for each part).</td></tr>
<tr><td>Quantity Consumed</td><td>This shows the number of each part consumed at each site for any given week. Each entry is tied to a part number and a site.

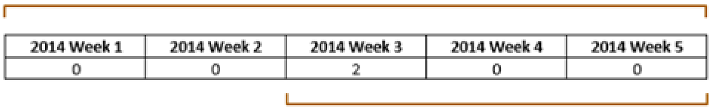


<table>
<tr><th>str_org</th><th>trl_part</th><th>yearweek</th><th>qty</th></tr>
<tr><td>PHX6</td><td>10001</td><td>201401</td><td>78</td></tr>
<tr><td>IND1</td><td>10042</td><td>201401</td><td>5</td></tr>
<tr><td>ABE3</td><td>10043</td><td>201401</td><td>1</td></tr>
<tr><td>BNA3</td><td>10044</td><td>201401</td><td>1</td></tr>
<tr><td>RIC1</td><td>10085</td><td>201401</td><td>1</td></tr>
<tr><td>PHX6</td><td>10086</td><td>201401</td><td>2</td></tr>
<tr><td>PHX3</td><td>10097</td><td>201401</td><td>3</td></tr>
<tr><td>PHX6</td><td>10097</td><td>201401</td><td>6</td></tr>
<tr><td>IND1</td><td>10106</td><td>201401</td><td>10</td></tr>
</table>

Figure 4.1: Example of part consumption in various sites in 1st week of 2014</td></tr>
</table>

<table>
<tr><td>First Sighted Date</td><td>This is the date the part was first sighted within the system and is necessary to accurately calculate the weekly average demand for a part.

$$\mathit{Weekly\ Consumption} = \frac{\sum \mathit{Weekly\ Consumption}}{\mathit{Weeks\ in\ System}}$$



<table>
<tr><th>2014 Week 1</th><th>2014 Week 2</th><th>2014 Week 3</th><th>2014 Week 4</th><th>2014 Week 5</th></tr>
<tr><td>0</td><td>0</td><td>2</td><td>0</td><td>0</td></tr>
</table>

For example, assume the table above shows the weekly demand for a part. If the demand is calculated as is, its demand would be 0.4 parts/week. However, if it is known that the part entered the system in Week 3, its average demand would be 0.66/week.</td></tr>
<tr><td>Store Min/Max Level and Current Quantity in Store</td><td>This shows the current min/max levels for every part in each site. This, together with the current quantity, is required for benchmarking the efficacy of the proposed inventory policy/pooling methodologies.</td></tr>
<tr><td>Lead Time / Avg. Price</td><td>The average lead time and price is also available for any given part at a specific site. This is again required for quantifying proposed changes, but also required as a key parameter in applying the (R,Q) model.</td></tr>
<tr><td>Criticality</td><td>The criticality of a part ranges from 1 to 3, and indicates how important a part is to a site (1 being the highest) and is similar to the A-B-C classification that Silver proposes for part prioritization.</td></tr>
</table>

Table 4.1: Parameters from SQL

4.2. Estimating Missing Data

Inventory models require specific inputs to generate results. For example, the (R,Q) model requires several parameters such as:

Usage data: Mean and standard deviation build a demand profile for the part.

Price: Affects EOQ and allows for quantification of savings

Lead Time: Dictates how much stock needs to be kept on hand (Cycle stock).

Criticality: Determines required service level for the part.

Any missing parameter would result in the model not being able to generate an optimal inventory stocking level for the part. Additionally, inaccuracies and missing data prevent the inventory model from being applies to all parts within the system. At present, only 57% of parts (in total monetary value) have sufficient data to allow for an implementation of an inventory model.

The following methods were employed to estimate the necessary parameters (where possible) and only parameters that could be estimated with a sufficient level of confidence were included within the study. This was able to improve the total parts included in the study from 57% to 67%, which still excluded approximately 33% of parts (in total monetary value). However, estimating parameters for the remaining parts could result in inaccurate values that would invalidate the entire model.

Although EAM provides each part in each site with its own parameters, a "common" parameter for each part has to be established for pooling purposes (i.e. a part must have one price within the system when pooled as opposed to having an individual price for each site). As the final goal of this project involves pooling, it makes sense to establish this common parameter and, if there is sufficient confidence in the "common" parameter, to apply that common parameter as the estimated data.

Criticality

Criticality values define the required service value for the inventory model. Due to the potential consequences of stockouts, if no parts within the network have an assigned criticality, the criticality for that part is assumed to be 1 (the highest criticality).

All parts take on the highest criticality assigned to the part across all sites. It is acknowledged that this may be overly conservative as sites with ten processing lines may have a low criticality for a part, whereas sites with only two lines may have the exact same part assigned a high criticality. However, due to the high service level requirements, this was chosen as the best compromise.

Lead Time

Lead times directly affect the inventory that has to be kept on hand due to its direct relationship with cycle stock. It is assumed that the bulk of the lead time comes from the manufacturer/supplier and not due to shipping. As such, for any given part, it is assumed that lead times will be fairly similar between sites.

All lead times of 0 are assumed to be missing data and will require estimation. An average is taken of all parts with non-zero lead time data. However, once an average is found, any parts with a lead time of +/- 1.282 SDs away from the mean is removed from the data (to eliminate potential outlier data), and a new average calculated. This mean is taken as the estimated lead time for that part and applied to all parts across all sites. In the event that there is no way of estimating a lead time for the part, it is assumed to be 4 weeks – the lead time for the vast majority of parts as shown from the distribution of parts below.

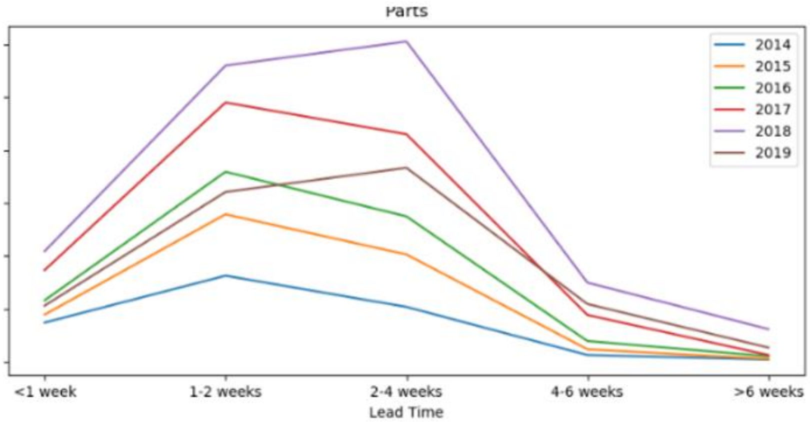


Figure 4.2: Lead Time Distribution for Parts

Price

Prices affect the optimal ordering quantity (a cheaper price would result in a higher EOQ) and are vital for quantifying results. As parts are secured from different suppliers, parts will have different prices within the system. Additionally, as these prices are input manually, they are also prone to errors (e.g. a $400 part may be input as $40,000) which could greatly skew recommendations. This was the area with the highest variation and risk of inaccuracies.

A common price, established through determining a mean/median price, helps to mitigate many of these problems. As before, it is assumed that although there are variations, the price for any given part is relatively similar across sites. Missing prices were estimated in three ways. These are presented based below in decreasing order of priority (i.e. the model will try to estimate prices using method 1 first when it encounters a missing price):

1) Preferred Supplier's Price of Part (within EAM)
2) Average price of equivalent parts in other sites
3) Average price of that particular part class (e.g. roller)

Once all parts have an estimated price, the model identifies a common price for each part. As before, all prices of 0 are assumed to be missing data. At this point, the model then identifies the mean and median price for each part based on the remaining price data.

The model then overwrites each part with the "common" price – either the mean or the median price for each part. These data points are presented below. Pre-consolidation refers to the sum of prices before the prices were overwritten and is adjusted to be $200 million for privacy reasons. The median and mean prices were also adjusted accordingly, keeping the ratios constant.

Configuration	Total Price of all Parts	Ratio to Pre-Consolidation
Pre-Consolidation	$200,000,000	1
Median	$206,990,387	1.035
Mean	$231,333,947	1.157

Table 4.2: Median vs Mean Price

In both cases, the median and mean both report higher prices than the pre-consolidation total. This suggests that many parts have under-reported prices within the system. An aggregation of the data shows many parts have a price of $1, which is not unexpected given that the recorded price within EAM used to be of no importance to the overall ordering process.

It is also worth noting that the mean prices are much higher than median prices, which suggest the presence of large outlier prices skewing the data upwards (as opposed to median prices were effectively eliminates larger outliers). Although the specific data cannot be shared, it is worth noting that the number of parts that had heavily under-reported prices were consistent across both the mean and median price consolidation methods.

In further support of price consolidation, there were a significant number of parts whose individual prices were significantly higher/lower than the mean/median price. For example, approximately 1,500 parts had a reported cost than was <10% of the mean cost, and approximately 400 parts had a reported cost that was 10,000% higher than the median cost. This highlights the variability in price and why consolidation is required for comparison purposes (i.e. It is not realistic for one site to buy a part for $210, whereas another site purchases the same part for $10).

Based on the findings from this analysis, the median price was used as the "standard" price. It was found that most parts were accurately reported (with most actual prices ranging between $\pm 10\%$ of the median), but outliers were typically more than an order of magnitude off. As such, the median was used as it remains true to the pre-adjustment value, but eliminates outlier values (e.g. $1 parts or overly expensive parts) that would skew results in mean-based reporting.

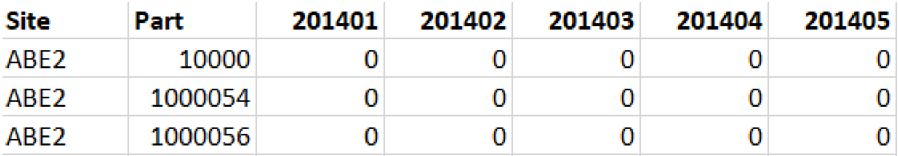
Usage Data

Usage data is exceptionally important as it provides a demand profile for each part and determines how much inventory needs to be kept on-hand. Due to the importance of this data, average demands and standard deviations are not estimated. If a part does not have usage data, it is excluded from the analysis.

4.3. Improving Data Accuracy – Returns and Outliers

In order to further improve the accuracy of the model, it was necessary to address issues that could skew the results. Trending of data points identified two areas of concern– extreme outliers and returned parts.

Outliers are especially concerning due to the very sparse demand data for most parts. Unexpectedly, most spares are unused from week-to-week and weekly demand for most part has a mode of zero.



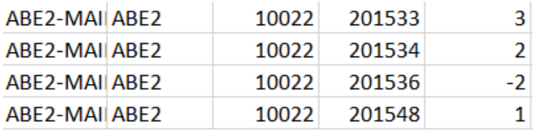
Site	Part	201401	201402	201403	201404	201405
ABE2	10000	0	0	0	0	0
ABE2	1000054	0	0	0	0	0
ABE2	1000056	0	0	0	0	0

Figure 4.3: Sample Week-By-Week Demand for 3 Parts

The sparse demand profiles exacerbate the impact of outliers in demand data, particularly when the part has not been on the shelves for very long. Additionally, this means that average weekly demand numbers do not necessarily reflect the actual weekly consumption required when parts are actually used.

Returns

The data management software allows for unused spare parts to be "returned" to the available spares after it has been checked out. For example, if three parts were taken out for a job, but only two were used, the last part can be returned upon completion of the job. Any returns to the system are input as negative uses which replenishes inventory levels for that part as shown below (item receivals are recorded separately, so there is no risk of confusion).



ABE2-MAI	ABE2	10022	201533	3
ABE2-MAI	ABE2	10022	201534	2
ABE2-MAI	ABE2	10022	201536	-2
ABE2-MAI	ABE2	10022	201548	1

Figure 4.4: Example of Negative Usage Data for Part 10022 in Site ABE2

Unfortunately, as seen above, these returns may not be identified until a few weeks later. If parts are checked out but returned in a different week, this will result in incorrect demand reporting. In order to address this, it was assumed that returns can only occur after a part is checked out. As such, all negative usages were assumed to be cancelled out by previous positive usages (starting with the most recently checked out part and working backwards). For example, in the example above, the two negative usages in the 36th week of 2015 would cancel out the two parts "used" in the 34th week.

In order to prevent incorrect "return" entries from corrupting the data, it is assumed that if a return is not fully removed after four weeks of "backtracking", there was a mistake in the quantity associated with the return, and any remaining amount to be returned is automatically deleted (with no other usages being deleted). This is able to correct the data to produce the cost chart as shown to the right below.

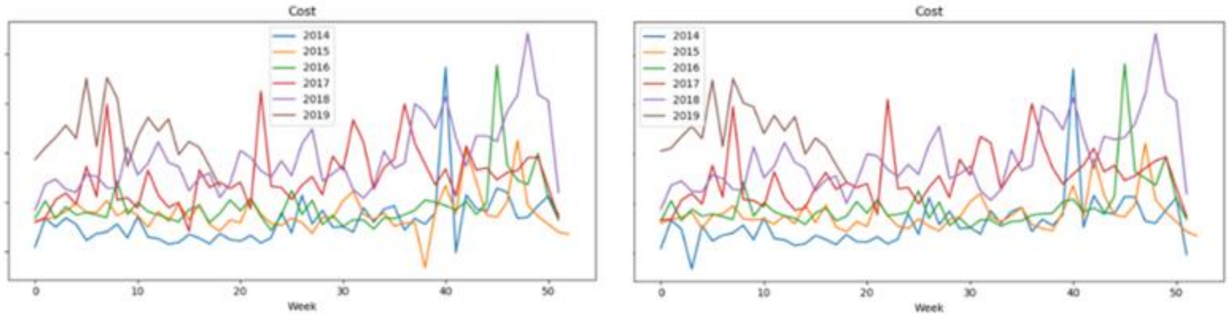


Figure 4.5: Consumption of Parts by Cost (Original vs Smoothed)

Although details on axis ticks were removed for data privacy reasons, it can be assumed that the lowest tick on the y axis corresponds to $0 total cost. Notice that the two negative spikes in ~weeks 38 and 41 have now been removed. Although this resulted in a negative spike in ~week 3, that negative spike was later found to be due to a negative price associated with a part in EAM (which was then removed from the dataset). The next section addresses how these outliers were identified and addressed.

Outliers

As highlighted by the previous section on estimating parameters, outliers present a significant risk to the accuracy of the model. This is exacerbated by the sparse demand data for most parts. Due to the size of the data (>1M data points), it was not feasible to investigate each data point. However, large outliers were identified by plotting the weekly consumption for parts and identifying large spikes in the data.

Plotting the weekly cost of parts consumption provided a way to identify abnormally high demand and high prices within the data as shown in the plots below. The plot on the left shows the original data. Each extreme point was individually and EAM admins were consulted when cases could not be immediately ruled out. For example, in the chart to the left, ~Week 40 of 2014 and ~Week 45 of 2016 show large spikes that warranted further attention. All unrealistic data points were removed which resulted in a much smoother chart as shown in the plot to the right, although spikes still do exist as part of normal operations.

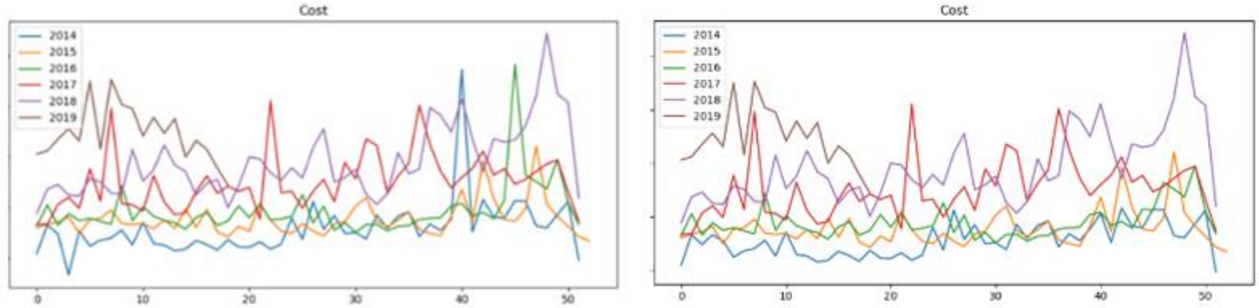


Figure 4.6: Consumption of Parts by Cost (Before / After Adjustment)

Additionally, plots were also created for total parts consumed per week. This ensured that parts with high demand (but inaccurately low prices) were not missed. For example, the plot below helped to identify a large spike in demand on the 30th week of 2017 that would have been missed by evaluating cost alone.

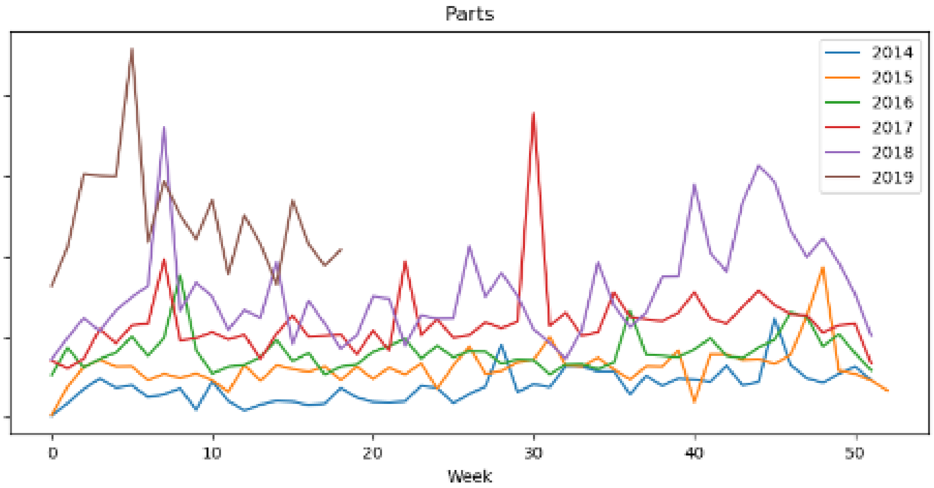


Figure 4.7:Weekly Consumption Parts (by Number of Parts)

4.4. Data Analysis

The above sections highlighted the data that was extracted from EAM, how missing data was estimated, and how outliers were identified and removed from the dataset. As mentioned in Section 4.1.1, the weekly demand statistics (average demand and standard deviation) could be calculated based on the consumption numbers and the known “first sighted” date of the part.

The dataset provides all the parameters needed to breakdown the data and identify key areas for improvement and answer high level questions that would impact the overall strategy regarding inventory.

4.4.1. Aggregation of Data

Although the analyses yielded useful data for prioritizing the inventory optimization methodology, they cannot be shared due to information that is confidential to Amazon. A small number of the data trends are presented below as examples to provide guidance for similar studies.

For example, Figure 4.8 below shows how the total cost and number of parts vary as the price of the parts increase. In this example, the data shows that the most expensive parts make up the largest proportion of cost, but make up a small proportion of the number of parts kept on-hand. Consequently, a larger focus was placed on ensuring that the more expensive marks were prioritized in the study.

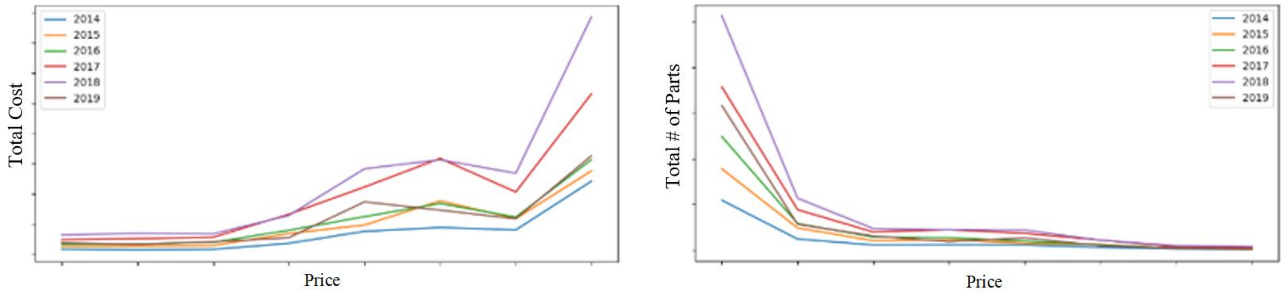


Figure 4.8: Price vs Total Cost / # of Parts

4.4.2. Managing Peak Seasons

Amazon faces two peak seasons throughout the year – Prime Day and the Christmas holidays. These two peak periods typically result in higher traffic through FCs. One concern was the need to ensure higher reliability during this period, and that there may be an uptick in parts usage during and prior to peak seasons (whether due to preventative maintenance, or higher loads causing faster spare consumption). If there is a significant increase in part consumption, a separate inventory model (with higher quantities of spares) may be required for certain periods of the year to ensure an adequate service level during peak periods.

This hypothesis was disproved by plotting an “adjusted” consumption chart that scales the parts consumption in each year so that the weekly demands can be better compared between years. Figure 4.9 suggests appears to be a peak in late February/early March, but when this peak was further examined in an unadjusted chart, the spike was found to be not significantly different enough to warrant a separate inventory policy.

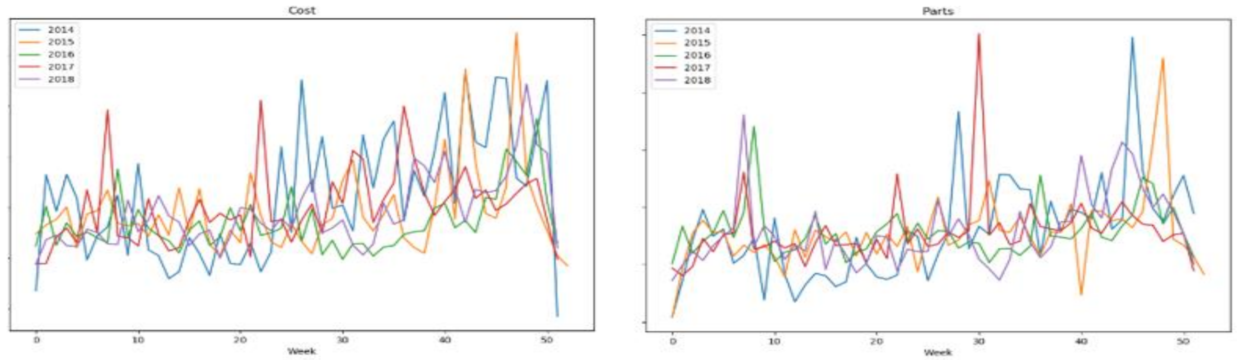


Figure 4.9: Scale Adjusted Consumption Plots

The expectation is that any gains from implementing two separate models would be small enough that they are offset by the complications of managing two inventory models. Based on this, this paper did not further investigate the applications of having different inventory policies for different time periods as there was insufficient evidence that peak periods exhibited higher than average demand.

4.5. Summary

In order for established inventory models, such as the (R,Q) model, to be implemented, a large amount of data is required. In particular, the parameters of criticality, price, lead time and usage data are especially important. Although EAM makes a large amount of this data available, there were still instances where that data was not available. Efforts were made to estimate these parameters where possible, but is not always possible. For example, any attempts at estimating usage data would only make the data less accurate.

As data accuracy is exceptionally important in ensuring that the proposed inventory policy is fit for use, data was aggregated to identify any potential outliers which could then be manually verified and eliminated. Finally, the aggregation of data also allowed for big picture analysis of the data, and identified that demand for spare parts was largely random and did not follow any seasonal trends. This was an important insight as it meant that a separate inventory policy did not have to be developed for Amazon’s two peak periods throughout the year.

5. Inventory Model Development

Amazon's current inventory management system does not make full use of the data that is available. As such, a single-stage inventory model is still likely to out-perform Amazon's current system. The proposed (R,Q) inventory model was chosen on the assumption that Amazon's supply chain involves an extended time horizon (i.e. multi-period model) and that parts will be reordered on a regular basis such that a continuous review period applies.

This chapter details how the (R,Q) model is applied within the Amazon context. Additionally, it goes into further detail regarding the background of the (R,Q) model by exploring its two building blocks - the Newsvendor and EOQ models.

5.1. Newsvendor Model

The newsvendor model is a single-period inventory model that ensures enough inventory is held to meet a required service level over a certain period of time (typically the lead time of the part). At a high level, the newsvendor model is made up of Cycle Stock (the expected amount of inventory required) and Safety Stock (the additional inventory carries to account for variations in demand).

The cycle stock and safety stock levels are a function of the various parameters that were determined in Chapter 4 and are shown below.

$$Cycle\ Stock: f(Average\ Demand, Lead\ Time)$$

$$Safety\ Stock: f(Standard\ Deviation\ of\ Demand, Service\ Level, Lead\ Time, Distribution\ Type)$$

The consumption of spare parts may be hard to predict, but the statistical parameters determined before allow for the random variable to be modelled through a probability density function as shown in Figure 5.1 below. The cycle stock is represented by cumulative distribution to the left of the mean (μ). The safety stock is some additional stock held on hand to account for the potential demand variation to the right of the mean. This serves as a buffer for higher than expected demand, and the specific amount of inventory held as safety stock will depend on the required service level.

The variable ‘k’ below represents the number of standard deviations from the mean required to hit the required service level. For example, a 95% service level will have a ‘k’ value of 1.65 (which corresponds to a cumulative distribution probability of 95%). The variable q* refers to the specific point on the distribution that results in the desired service level / cumulative probability.

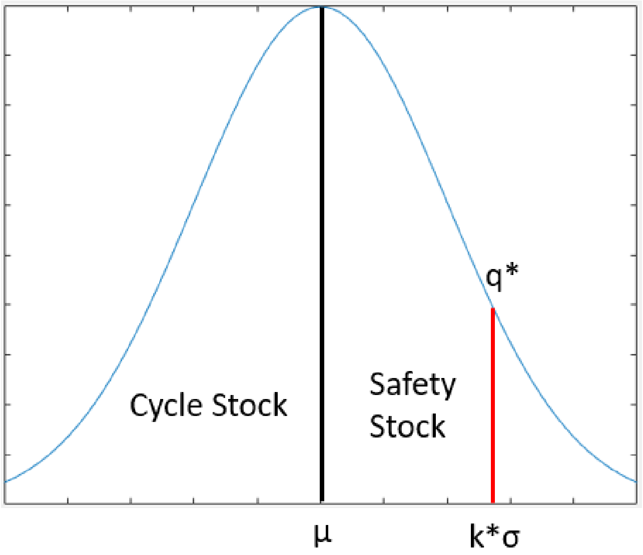


Figure 5.1:Newsvendor Diagram

A traditional newsvendor model will determine the optimal ‘k’ value through the overage and underage model described in Chapter 2. However, due to Amazon’s strict MHE uptime requirements, the underage costs are considered to be significantly higher than the underage costs, and are not easily quantified.

Although Amazon would prefer a service level of 100%, it should be noted that this is statistically impossible, and would require an infinite number of parts. Instead, the required service levels were manually selected based on the parts’ criticalities. It should be acknowledged that this was a deliberate decision that was made to optimize for uptime, but may come at a potentially higher cost (than if the overage/underage method were used).

Criticality	Service Level	k Score
1	99.5%	2.58
2	99%	2.33
3	95%	1.65

Table 5.1: Criticality to Service Level Mapping

Based on this criticality, the amount of inventory for each part required on-hand can be calculated by applying the formula below based on each part's individual parameters.

$$Inventory\ Required = Cycle\ Stock + Safety\ Stock$$

$$Inventory\ Required\ [Parts] = \left(\mu \left[\frac{Parts}{Time}\right] * Lead\ Time\ [Time]\right) + \left(k * \sigma \left[\frac{Parts}{Time}\right] * \sqrt{Lead\ Time}\ \ [Time]\right)$$

As evidenced above, the Newsvendor model requires an underlying distribution to build its recommended inventory (otherwise 'k' and σ do not exist). Although Figure 5.1 above uses a normal distribution for illustrative purposes, the stochastic nature of spare parts consumption may not follow a normal distribution.

5.1.1. Potential Distribution Types

There are a number of distribution types typically used for Newsvendor models. Examples include the Normal distribution, the Poisson distribution, and the Negative Binomial distribution (but is not limited to these). The following sections details the benefits and complications that arise from each distribution type to establish a baseline for selecting a distribution type.

Normal Distribution

The Normal distribution is one of the most preferred distribution types due to its versatility and ease of application. The cumulative distribution for Normal distributions is given by

$$q^* = \ \mu + k\sigma$$

Where k represents the number of standard deviations away from the mean required to achieve the desired cumulative probability.

The Normal distribution is a continuous distribution that is bounded by [-∞,∞]. This results in several problems when applying the Normal distribution to inventory.

First, inventory is usually best modelled through discrete distributions (with count variables). However, this can be mitigated by rounding all partial inventories upwards (to be conservative). Additionally, Amazon FCs often use "partial" parts for certain consumables (e.g. length of conveyor belt) which discrete distributions will not be able to adequately deal with.

The second point of consideration is that inventory cannot be negative (i.e. it should be bounded by [0,∞]). However, by definition, average demand (μ) for a part must always be greater than zero which means that negative values will always fall to the left of the mean. As Amazon's service levels (>95%) will always result in the required inventory being to the right of the mean, any negative values on the distribution curve can effectively be ignored.

Finally, although the Normal distribution may not perfectly map to the consumptions' distributions, the large number of data entries enables the Central Limit Theorem to hold, which supports the applications of a Normal distribution.

Poisson Distribution

The Poisson distribution is a discrete distribution that typically works well for count data (i.e. non-negative integer data {0, 1, 2…}). The Poisson distribution is described by a single parameter (λ) where λ > 0. By definition, the mean of a Poisson distribution is equal to its variance:

$$Y \sim P(\lambda = \mathrm{x}) = \frac{e^{-\lambda}\lambda^{\mathrm{x}}}{x!}$$

$$E(\mathrm{Y}) = Var(\mathrm{Y})$$

Additionally, the Poisson distribution models the sparse demand profile of spare parts quite well as it is very right-skewed with many data points at 0. Given that Poisson distributions are discrete distributions, the cumulative probability function is:

$$P(X \leq k) = \sum_{i=0}^{k} \frac{e^{-\lambda}\lambda^{\mathrm{i}}}{i!}$$

As such, q*, for a given service level (α) is such that:

$$\sum_{i=0}^{q^*-1} \frac{e^{-\lambda}\lambda^{\mathrm{i}}}{i!} \leq \alpha \leq \sum_{i=0}^{q^*} \frac{e^{-\lambda}\lambda^{\mathrm{i}}}{i!}$$

The Poisson is an especially good approximation when the data is under-dispersed (when the mean is less than, or equal to, the variance (i.e. (E(Y) ≤ Var(Y))). When the data is over-dispersed (i.e. (E(Y) > Var(Y)), the Negative Binomial Distribution becomes a better approximation. Based on the data available, there is a fairly even mix of over-dispersed and under-dispersed data, which makes it difficult to select between the two.

	Count (n = 100,000+)
E(Y) < Var(Y)	27.2%
E(Y) = Var(Y)	13.3%
E(Y) > Var(Y)	59.5%

Table 5.2: Mean vs Variance Distribution of Dataset

Negative Binomial Distribution

The Negative Binomial Distribution is similar to the Poisson distribution, but works better with over-dispersed data. Unfortunately, the data is a fairly balanced mix of over-dispersed and under-dispersed data (which would be better served by a Poisson distribution). As such, there is no clear way of choosing one distribution over the other.

The Negative Binomial distribution is typically presented as follows:

$$X \sim NB(r, p)$$

Where r and p are the parameters that define the distribution. This thesis will not discuss the details of this distribution, but it should be noted that unlike the Normal and Poisson distributions, the parameters r and p are not defined by the statistical mean and variance of the parts' demands. As such, any application of the negative binomial distribution will require additional time to compute the required parameters.

Empirical Distribution (Non-Parametrical)

The Empirical distribution is a non-parametrical distribution and is the preferred distribution given the available data as it does not attempt to reduce the distribution into several parameters. Instead, it simply uses the real, historical data to create a distribution model. This maintains the integrity of the available data and avoids overfitting which could result from forcing a distribution.

However, due to the relative size of the dataset, the empirical distribution cannot accurately reach the required service level of 99.5%. For example, 5 years of weekly data is approximately 260 data points. This would simply ignore the highest 2-3 data points to reach 99.5% data coverage, which is problematic as most parts are not used that regularly. Ignoring the result in the highest 2-3 data points would result in the large number of zeros being over-indexed within the distribution, and may even report a demand average of 0.

As such, although the Empirical distribution is the preferred distribution for data fidelity, it cannot be applied to Amazon's FCs until more data points are available.

Best-Fit Distribution

Rather than applying a blanket distribution to cover all parts, it is also possible to use computational methods to identify the best-fit distribution for every single part. This would ensure that the distribution chosen to represent the demand profile of the part. However, since there are tens of thousands of parts, this is an exceptionally time-consuming process if the simulation has to be run multiple times.

Although a best-fit distribution would provide the best possible outcome, it may not be feasible in an operational environment where the simulation will have to recalculate min/max levels on a regular basis (e.g. when a new part is added, when suppliers change lead times etc.) when there are potentially long computational times involved.

5.1.2. Distribution Selection

In order to establish an initial baseline (and since time was not a constraint), best-fit distributions were identified for every part by running each part's demand profile through MATLAB. The table below shows how often a distribution type was selected as the best-fit distribution.

Distribution Type	Proportion of Parts
Normal	0.03%
Exponential	25.51%
Gamma	10.01%
Chi	15.95%
Exponential-Normal	48.51%

Table 5.3: Best-Fit Distribution Types

As highlighted in section 5.1.1, being able to apply the normal distribution is ideal due to the ease of identifying the necessary parameters required for its implementation. As seen from the table above, very few parts are optimally represented by the Normal distribution, and the exponential-normal distribution appears to dominant. However, this table does not account for how far off the normal distribution is from the optimal exponential-normal distribution.

In order to test the adequacy of the various distributions (Normal, Binomial and Best-Fit), an inventory policy was developed using the (R,Q) model (see chapter 5.3) with each of the above distributions applied to all parts through the Newsvendor model. The table below shows the ratios of the normal and binomial distributions against the best-fit distribution with respect to the total value of parts when the (R,Q) model is implemented (absolute values not provided for data privacy).

Underlying Distribution	Ratio
Best Fit	1
Normal	1.012
Binomial	0.853

Table 5.4: Performance of Distributions vs Best-Fit

As the goal of this exercise is not to minimize the total value, but rather to find a distribution that best approximates (and is equal to) the best-fit distribution, it becomes clear that the binomial distribution under-represents the overall data. Although the normal distribution may not be ideal for many parts, its performance is not too far from the best-fit distributions.

As the input data is already inherently inaccurate, a difference of ~1.2% is within the allowable tolerance. Accordingly, a Normal distribution is assumed for all parts to allow for quick computations that are still relatively accurate.

5.2. EOQ

As described in Chapter 2, the EOQ aims to minimize the total costs associated with purchasing parts (the holding cost and the ordering cost). The EOQ quantity (Q) is given by:

$$Q\ [Parts] = \sqrt{\frac{2 * Demand\ \left[\frac{Parts}{Time}\right] * \ Fixed\ Ordering\ Cost\ [\$]}{Holding\ Cost\ [\frac{\$}{Time * Part}]}}$$

As it is not possible to order partial quantities, the EOQ has to be a minimum of 1. This identifies the optimal quantity that should be ordered at any given time.

However, several scenarios within Amazon’s spares inventory make it possible to further refine this EOQ quantity by considering NPV savings. This is further described in chapter 5.2.1.

5.2.1. EOQ Optimizations

Low part prices could typically result in high EOQs due to low holding costs. For example, the EOQ formula may order 200 weeks' worth of inventory to minimize costs if fixed costs are significantly higher. However, this consumes capital and takes up space within the spares cage. As such, where the EOQ is significantly higher than the weekly demand, there is a potential to realize further NPV savings and free up space by limiting the amount of EOQ purchased at any point in time.

One way of achieving this is by setting a "cap" to EOQ purchase orders based on reorder quantities where "tolerance" is the maximum number of reorder quantities (R) that a site should hold at any given time.

$$New\ EOQ = \min(Q_{EOQ}, Tolerance * R)$$

This essentially "caps" the inventory at a specific multiple of reorder quantities in order to lower the inventory on hand at any given time. Although this is no longer an "optimal" solution, it helps to reduce inventory levels when excessively high quantities are purchased.

By definition of the EOQ, if the reorder quantity (Q) is reduced through this multiple, this would in higher shipping costs. The number of shipments per week for any given part is given by:

$$\frac{Demand}{Q} = Shipments\ per\ Week$$

If a fixed cost of $10 per shipment is assumed (more details in Chapter 6), the shipping costs per year for that part comes to:

$$Annual\ Shipping\ Costs = \frac{Fixed\ Cost * Weekly\ Demand * Weeks}{Q} = \frac{10 * \mu * 52}{Q}$$

Note that since weekly demand is fixed, the only variable that affects the shipping costs is Q. As such, the change in annual shipping costs (from having a Q that differs from EOQ) can be calculated through:

$$\frac{520\mu}{Q_{new}} - \frac{520\mu}{Q_{EOQ}}$$

$$= \frac{520\mu * Q_{EOQ}}{Q_{new}Q_{EOQ}} - \frac{520\mu * Q_{new}}{Q_{new}Q_{EOQ}}$$

$$= \frac{520\mu * (Q_{EOQ} - Q_{new})}{Q_{new}Q_{EOQ}}$$

The financial impact of implementing this constraint are presented in the table below, along with the total number of parts that are taken off the shelves. The financial impact is presented as the NPV of shipping costs (based on Amazon's cost of capital). Again, the data has been masked, but have been kept in the same relative orders of magnitude for the illustrative purposes.

Tolerance (in R)	Parts Off Shelf	Additional Shipping Cost (NPV)
4	500,000	$950,000
8	250,000	$200,000
12	175,000	$75,000
16	120,000	$35,000

Table 5.5: EOQ Constraints

As shown in Table 5.5 above, as the tolerance increases, fewer parts are taken off the shelves (as the maximum allowable inventories are higher), and fewer additional shipments are required to replenish inventory levels. Figure 5.2 below shows the rate at which these two parameters (how many parts can be removed from shelves vs shipping cost) change as the tolerance varies. Note that the magnitude of the two lines are not comparable, but simply show the difference in the rate of change.

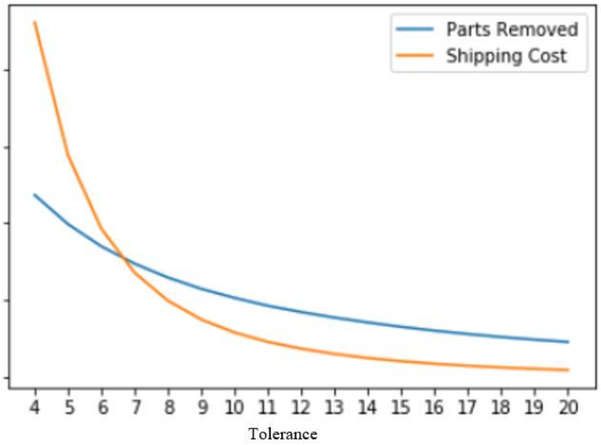


Figure 5.2: Increase in Shipping Costs due to EOQ Constraint

However, this neglects the additional benefit of postponing purchase of parts. If parts are purchased later in the year, it would free up present day cashflows, which has associated NPV savings. The formula below shows how such an NPV calculation would be done, where P is the price of the part.

$$\frac{Q_{new} * P}{1} + \frac{Q_{new} * P}{1 + \frac{Cost\ of\ Capital}{52} * Interval} + \frac{Q_{new} * P}{\left(1 + \frac{Cost\ of\ Capital}{52} * Interval\right)^2} \cdots$$

The interval between purchases is:

$$\frac{Q_{new}}{D}$$

Accordingly, the total number of periods in the NPV calculation is given by:

$$\frac{Q_{old}}{Q_{new}}$$

This provides a way of calculating the expected NPV savings alongside the additional shipping costs as demonstrated in Table 5.6 and Figure 5.3 below.

Tolerance (Weeks)	Parts Off Shelf	Additional Shipping Cost (NPV)	NPV Savings
4	473,213	$922,083	$130,433
8	258,510	$198,273	$47,342
12	168,793	$73,509	$23,310
16	120,453	$34,421	$13,201

Table 5.6: EOQ Constraints: Shipping Costs vs NPV Savings

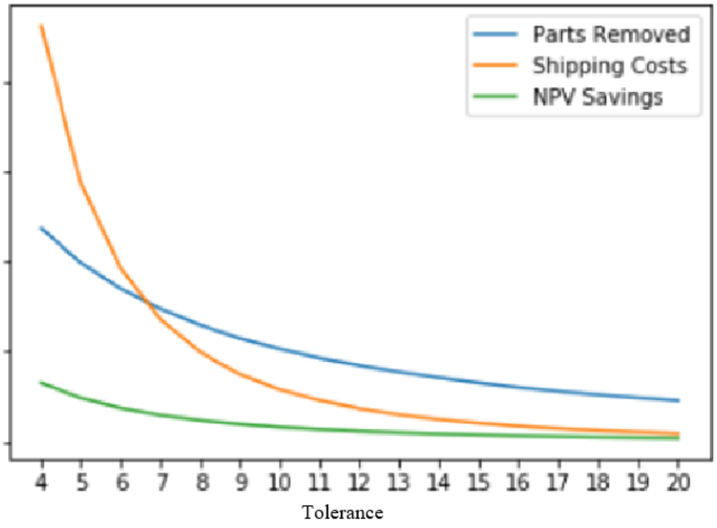


Figure 5.3:EOQ Constraints: Shipping Costs vs NPV Savings

There are two additional savings that result from postponement of purchasing parts. First, there is an immediate reduction in parts, which can be considered a once-off, immediate cost savings (remember that inventory on hand is the reorder amount (R) and the reorder quantity (Q), so any reduction in Q reduces total inventory within the network). Additionally, as there are fewer parts in the network, this reduces overall holding costs.

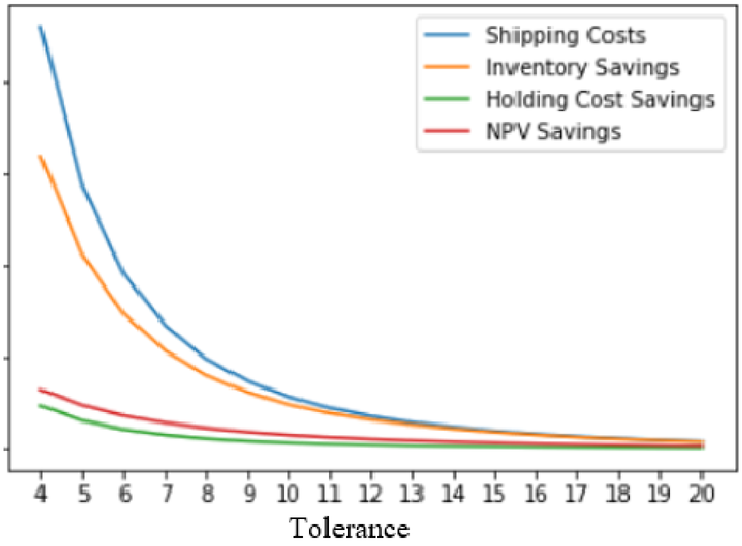


Figure 5.4:EOQ Constraints: Overall Cost Savings

All of these costs can be superimposed onto one graph as shown in Figure 5.5 below, such that negative costs represent savings. Accordingly, the optimal tolerance, from a cost perspective, to be approximately 7 reorder quantities for every part.

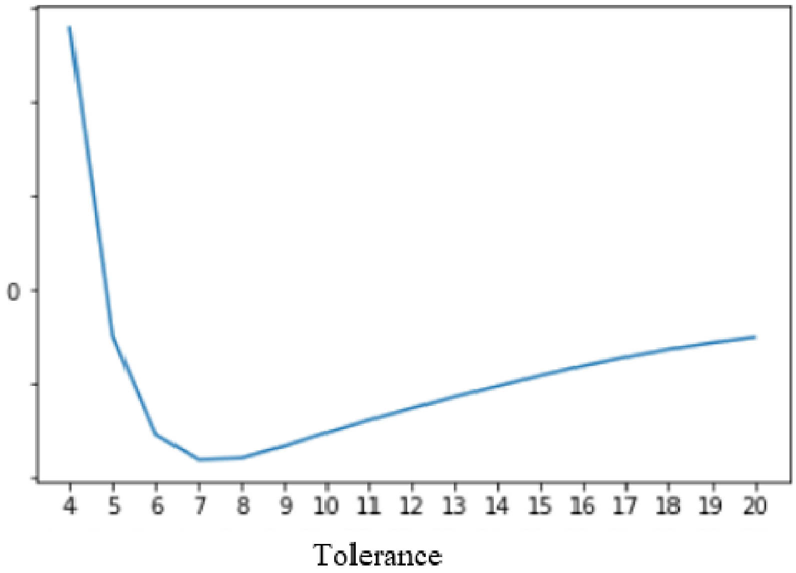


Figure 5.5: EOQ Constraints (Overall Savings)

Naturally, every part will have a different tolerance. If each part were to have its optimal tolerance identified and applied as opposed to the blanket tolerance of 7 reorder quantities, this would result in an additional 3% in savings. The added complexity of identifying the optimal tolerance and implementing it for hundreds of thousands of parts on a regular basis (as demand profiles change) was considered to be too time-consuming for practical purposes for 3% savings.

Accordingly, the tolerance for every part was taken to be 7 reorder quantities, and was only applied if it would result in net savings for that particular part.

$$New\ EOQ = \min(Q_{EOQ}, 7 * R)$$

5.3. (R, Q) Inventory Policy

The (R, Q) inventory policy builds upon the two building blocks of the Newsvendor model and the EOQ. As detailed in Chapter 2, the Reorder Point (R) is based upon the Newsvendor model to ensure enough inventory is kept on-hand to cover demand and its associated fluctuations, and the Reorder Quantity (Q) is based on the EOQ to minimize ordering costs.

$$R = \min\left(1, \mu * Lead\ Time + \ \sigma * k * \sqrt{Lead\ Time}\right)$$

$$EOQ = \min\left(7 * \mathrm{R}, \min\left(1, \sqrt{\frac{2 * \mu * Fixed\ Cost\ per\ Order}{Weekly\ Holding\ Cost}}\right)\right)$$

Note that both R and Q have a minimum of 1. The reorder point needs to be a non-zero value due to Amazon’s desire to not enact a pull model for spare parts (to ensure maximum uptime). Additionally, the reorder quantity has to be non-zero to ensure that actual orders are placed.

As found before, it is assumed that the normal distribution applies to every part. This greatly simplifies the calculation for k as the normal distribution is exceptionally well defined. Although it would be ideal to achieve a 100% service level for each part, this is not feasible as it would require an infinite number of parts to ensure perpetual uptime (statistically speaking).

As such, the criticality of each part within the FC determines its required service level, and parts are stocked accordingly to meet that requirement using the Newsvendor model.

Criticality	Service Level	k Score
1	99.5%	2.58
2	99%	2.33
3	95%	1.65

Table 5.7: Criticality to Service Level

The (R,Q) model then forms the baseline inventory policy upon which the rest of this study can be built. It is worth noting that, this (R, Q) inventory policy alone could be one phase of reducing inventory across the FCs as it will optimize all inventory levels.

If the (R,Q) model proposes a decrease in inventory, this would serve our desired outcome. However, if the (R,Q) model proposes an increase in inventory (from current stocking levels), that would imply that the site currently does not stock enough parts to meet the desired service level for that part.

The table below shows the results of how the proposed (R,Q) model performs against Amazon's current spares inventory (Again, numbers have been masked, but kept around similar orders of magnitude).

Required Change to Inventory	**# SKUs**	**Change to Quantity of Parts**	**Change to Total Value**
Decrease (Currently Over-Ordering)	~50,000	~500,000 parts	~$50,000,000 Reduction
No Change	~50,000	N/A	N/A
Increase Levels (Currently Under-Ordering)	~60,000	~1,100,000 parts	~$35,000,000 Increase

Table 5.8: Differences in Inventory Levels (Current Min/Max vs Proposed (R,Q) Model)

Based on the table above, it appears that low-value parts were typically understocked, whereas FCs typically held a larger than required amount of high valued parts. Out of the entire NAFC network, only 1 site had >50% of its SKUs stocked at the optimal levels proposed by the (R,Q) model. This further highlights the arbitrary nature of Min/Max numbers and the need to move towards data-based inventory policies.

5.4. Conclusion

The (R,Q) model provides a way in which FCs can apply an inventory model to their spare parts. Successful implementation of the (R,Q) model could reduce total inventory costs across the entire FC network by approximately 10%. Although there are several ways in which this can be further refined, for example, by setting a cap on the amount of spare parts that can be purchased through the EOQ, there remains few other opportunities to further reduce inventory levels without affecting day-to-day operations unless significant changes are made to suppliers and/or current maintenance programs.

A multi-echelon supply chain presents an opportunity to further reduce inventory levels by a further 15%. The methodology for a multi-echelon supply chain is presented in the following chapter.

6. Multi-Echelon Inventory

Multi-echelon inventory systems provide a way of further reducing inventory levels across the network by pooling sites together and reducing the overall variation in the system demand. This chapter explores the concept of multi-echelon systems and how it can be applied to Amazon's FCs. In particular, this chapter develops methods to optimize the number and locations of centralized hubs required, as well as the frequency of delivery. The chapter closes by performing several sensitivity analyses to determine how the various levers can impact the efficacy of a multi-echelon system.

6.1. Multi-Echelon Systems

Once inventory levels have been optimized through an inventory policy, there are limited opportunities to further improve the current system due to lead time and service level constraints. A multi-echelon system (also known as nodal warehousing) introduces centralized "hubs" that sit in between the suppliers and the FCs, and serve as intermediate staging points that support FCs. In this system, the hubs hold their own inventory and order directly from suppliers, while FCs replenish used parts directly from the hubs.

This system decreases inventory in two ways. First, FCs will receive parts from the hub in a matter of days (as opposed to weeks). This allows FCs to significantly reduce their inventory levels. Second, FCs no longer have to hold large amounts of "safety" stock to account for larger than expected surges in demand as nodal warehousing allows FCs to pool their usages. As more FCs are pooled together, usage peaks at one site are more likely to be balanced out by lower consumption at a different site (i.e. The more FCs a hub supports, the more it is protected from demand variability). This aggregation of demand variability shields the network from demand fluctuations and allows for lower inventory levels across the network.

It is worth noting that not all parts should be stocked at the hub – FCs should still order some directly from suppliers. Although pooling inventory decreases inventory levels across the network, it adds an additional shipping leg into the supply chain. As such, inventory should only be pooled if the reduction in inventory from pooling makes up for the increased shipping costs. For example, if there is insufficient variability in a part’s demand profile, or if a part is not widely used, the additional shipping costs from hub to FC may outweigh the reduction in inventory levels from pooling.

Intuitively, regardless of whether the supply chain is single stage or multi-echelon, the demand from FCs does not change. Instead, any reduction in inventory is less capital that is held in inventory (and can be considered as an immediate saving). If the increase in shipping cost from implementing a multi-echelon network does not exceed this saving, then the multi-echelon network is worth pursuing.

6.2. Facility Location Analysis

The goal of this phase of the project was to identify the optimal number of hubs, and where those hubs should be located. The process of optimally placing a facility to minimize costs is a well-researched branch of operations research known as location analysis (or facility location problems). For the purposes of this study, there are no constraints and the facilities can be placed in any location. It is assumed that all facilities deployed for pooling inventory does not have a maximum capacity, making this an uncapacitated facility location problem.

The main consideration is determining hub quantity/location is the balance between additional shipping costs versus lower inventory volumes. In order to determine the optimal hub parameters, virtual facilities were created in various pre-determined locations to test their efficacy in lowering inventory across the network. This would allow for calculations of potential savings across all identified locations. For simplicity, it is assumed that shipping cost is standardized based on distance.

However, there are also other factors to consider such as real estate costs and labor costs. Fortunately, these costs do not have to be considered immediately as they do not influence the shipping vs. inventory reduction calculation. Instead, they can simply be added on at the end after total savings have been calculated.

As shipping accounts for a significant proportion of the operational cost for a multi-echelon system, the ideal site(s) would minimize the distance between the warehouse(s) and the FCs:

$$\min\left(\sum |Distance\ from\ Hub\ to\ FC|\right)$$

Although this does not guarantee optimal savings, it will help to narrow down the list of sites to consider.

6.2.1. Potential Hub Locations

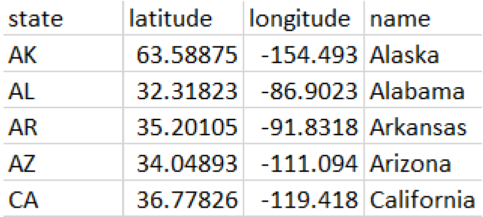
A wide range of potential hub locations were chosen with the goal of evaluating how location will impact overall shipping costs. Two methods were developed for determining potential hub locations:

1) Randomized coordinates within each state;
2) Nodes with a high concentration of FCs (e.g. 3+ FCs within 50 miles of each other).

In order to ensure that the analysis was sufficiently thorough, all states were included in the study, even if they did not make intuitive sense (e.g. a hub in Alaska would be too far removed to serve as a proper hub, but was included in the study regardless).

As the goal of this initial step was to identify general location trends, hubs could be placed at a relatively central point in each state. Although these centralized locations may not be the ideal position for the hub, they would provide the necessary benchmarking data to guide for further analysis. Detailed selection of hub location would be possible once optimal locations (i.e. states) were identified.

As FC locations are available in longitude and latitude, hub locations were also similarly identified by their longitude and latitude (taken from https://developers.google.com/public-data/docs/canonical/states_csv). The table below shows examples of hub locations in five states.



state	latitude	longitude	name
AK	63.58875	-154.493	Alaska
AL	32.31823	-86.9023	Alabama
AR	35.20105	-91.8318	Arkansas
AZ	34.04893	-111.094	Arizona
CA	36.77826	-119.418	California

Figure 6.1: Selection of Potential Hub Locations

6.2.2. Distance Calculations

In order to calculate the shipping cost between the hub and FCs, the distance between the potential hub(s) and FCs were needed. As the longitudes and latitudes for potential hub locations were found in Section 6.2.1, they could be mapped to the known FC locations. When testing network configurations that involved multiple hubs, FCs were supported by the hub that was closest to it.

There are two available methods for computing distances: Vincenty's formula and the Haversine formula. Vincenty is generally regarded as more accurate at estimating distances when compared to Haversine as it accounts for the Earth not being perfectly spherical. However, this accuracy results in Vincenty being more computationally intensive. Since precise measures are not yet required at this stage, the model calculates distance between hubs and FCs through the Haversine formula, which provides the straight-line distance between two sets of longitudes and latitudes.

$$d_{12} = Distance\ between\ Points\ 1\ and\ 2$$

$$d_{12} = 2 * Earth\ Radius * \arcsin\left(\sqrt{sin^2\left(\frac{lat_2 - lat_1}{2}\right) + cos(lat_1)\, cos(lat_2)\, sin^2\left(\frac{long_2 - long_1}{2}\right)}\right)$$

As most roads aren't straight, a 25% multiplier is further added onto this distance to more accurately represent the road distance between sites. This 25% multiplier was found to be a reasonable ballpark after testing several distances and comparing it to known road distances. Applying this formula to all hub/FC combinations then develops a full set of distances between all hubs and FCs.

6.2.3. Zone Modelling

Once distances between proposed hub locations and FCs are known, it is possible to calculate expected shipping costs. It is assumed that third-party shipping (e.g. USPS) is used as it is immediately available and has minimal barriers to implementation. Additionally, for purposes of a feasibility study, it provides an upper-bound estimate for shipping costs (i.e. other shipping methods will only be used if they result in lower costs). Standard ground shipping was used as a baseline as it provides the optimal balance between lead time (lower inventory levels) and shipping costs.

Third-party shipping typically calculates shipping rates based on weight and zones (distance bandings) as shown in the table below. Essentially, as long as two FCs are within the same zone from a hub, the shipping price from the hub to the FCs is the same even if one FC is slightly further away. Although Amazon's third-party shipping rates have been masked for the purposes of this thesis, the table below of retail rates provide an upper bound estimate of shipping prices between two sites based on the number of zones between them.

Weight	Zone 1-2	Zone 3	Zone 4	Zone 5
10lbs	$10.84	$11.51	$12.60	$13.76
20lbs	$13.08	$14.75	$15.06	$18.19
30lbs	$15.86	$18.61	$20.64	$24.49
40lbs	$18.07	$22.05	$25.01	$30.15

Table 6.1: Example of Third-Party Shipping Rates

Although shipping costs vary by weight, they have relatively consistent pricing ratios between zones (independent of weight). This paper assumes that packages are approximately 10lbs (the most common weight of parts shipments).

Based on a standardized package weight, the table below shows the shipping price for a single package from hub to FC based on the zones travelled. The shipping price shown in Table 6.2 is an approximate average of several third-party shipping prices for a 10lbs package. The price differential is also provided as a "zone multiplier" which provides comparative ratios between shipping prices for various zones. In this instance, the zone multiplier is found by dividing each zone's shipping price by the base shipping price of

$10. This is used over the shipping price as it provides more flexibility in price adjustments (as will be seen in Section 6.4.2 on shipping prices).

Zones Travelled	Distance Travelled [mi]	Shipping Price [$]	Zone Multiplier
1-2	0 – 150	10	1
3	150 – 300	11.50	1.15
4	300 – 600	12.50	1.25
5	600 – 1000	15.00	1.5
6	1000 – 1400	18.00	1.8
7	1400 – 1800	21.00	2.1
8	1800+	24.00	2.4

Table 6.2: Shipping Zone Multipliers

Due to zone pricing, the optimization function should no longer be for minimum distance, but rather for the lowest zone multiplier between the selected hub(s)s to the FCs. Since the zone multiplier is reflective of the distances between sites, the new objective function can be given by:

$$\min\left(\sum |Total\ Zone\ Multiplier\ |\right)$$

6.3. (S-1, S) vs (R, Q) Inventory Models in Multi-Echelon Networks

The (R,Q) model determines the minimum number of parts to stock (through the Reorder Point), and the optimal ordering quantity (Q) that minimizes overall costs associated with placing an order. This model is especially important in the current system as parts are ordered from multiple suppliers. However, nodal warehousing introduces a central hub that holds parts from every supplier, which means that FCs can receive all of its required parts in one shipment. This means that minimizing costs for an individual shipment is no longer the most efficient method at an FC level.

Nodal warehousing could make it cheaper to stock parts using an (R,Q) model where Q = 1 (i.e. FCs reorder parts as soon as they are consumed). More specifically, this is a return to the Min/Max inventory model where the minimum is the Reorder Point, and FC reorder every day to stay one part above the minimum. This is known as the (S-1, S) model, or the base stock model. In the (S-1, S) model, inventory levels are lower as FCs only reorder what is required, as opposed to purchasing Q and running down that inventory over time. As inventory levels are given by R+Q, a smaller Q would result in lower inventory levels.

Since the model deviates from the optimal ordering quantity, there may be an increase in total shipping costs. However, most sites will likely require daily shipments from the hub to restock any parts consumed on the day. Notice in the table above that sending a 20lbs package is significantly cheaper than sending two separate 10lbs packages. As such, FCs also benefit from pooled shipping when using a multi-echelon supply chain.

This model is only feasible within a nodal network due to short lead times between the hub and FCs, and the fact that hubs will restock FCs on an almost-daily basis. The analyses in the following sections will explore how the multi-echelon supply chain performs using both the (R,Q) and (S-1, S) inventory models in FCs. Hubs always operate under an (R,Q) policy as they still order directly from suppliers.

6.4. Hub Selection

Although nodal warehouses can serve as an intermediate point for inventory, not all inventory should be stocked centrally. As shown in Figure 6.2 below, a nodal network actually increases the total lead time between the supplier to the FCs due to the additional shipping leg between the hub and the FCs. The inventory reduction from a nodal warehousing strategy comes from the reduction in variability by combining multiple sites, which allows for a reduction of network safety stock.

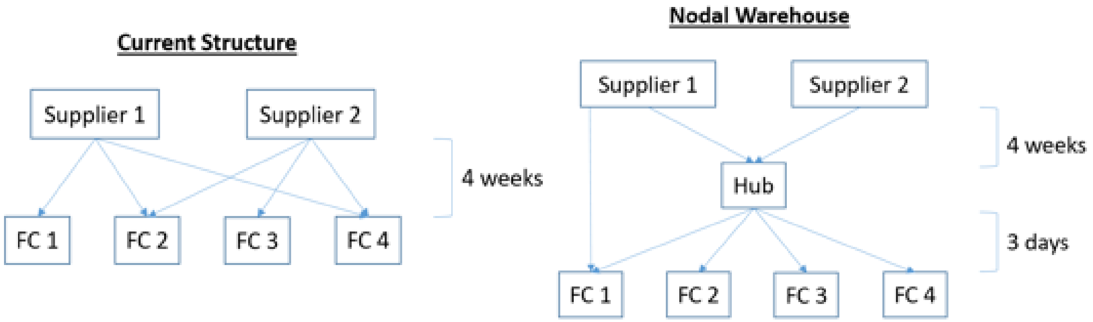


Figure 6.2: Current Structure vs Multi-Echelon Network

Due to this increase in lead time, if a hub does not actually reduce total inventory held within the network for a particular part, it would be better for that part to held directly at the FC (which then orders those parts directly from the supplier rather than the hub). Not only does this reduce logistical complexity, but it also removes one leg of shipping (and cost) from the supply chain. As such, there has to be a method to determine if a part should be stocked at the FC or at the hub before the savings from a hub can be calculated.

6.4.1. Inventory Placement

The inventory level of a hub captures the entire system downstream of it. As such, a hub's inventory is the combination of its inventory on-hand, the inventory in transit to FCs, and the inventory of all the FCs that it supports.

$$I_{hub} = I_{hub,on-hand} + I_{transit} + \sum_{j=1}^{N} I_{fc_j} \qquad (N = \#\ of\ Supported\ FCs)$$

Accordingly, for every part, the model calculates the amount of parts required if it were stocked centrally, and compare that to if it were stocked at an FC-level. If the inventory reduction from central stocking makes up for the increased shipping cost, then the part is stocked centrally. If nodal warehousing does not lower net inventory, or if the savings does not make up for increased shipping cost, then the part is not pooled.

First, each FC is assigned to a hub based on proximity (if there is only one hub, all FCs are assigned to the same hub). This allows populates a list of potential parts that the hub will need to stock (i.e. In a two-hub configuration, if none of the sites supported by one hub use part “10000”, that hub will not need to stock part “10000”).

In order to calculate the required inventory level for each part, the combined demand for the part within the hub’s network must be known. As before, it is assumed that the central limit theorem holds and that the distribution is normal. The average demand in the hub is the sum of the downstream demands. To put this in plain terms – if one site uses 3 parts/week, and another site uses 4 parts/week, the combined usage is, on average, 7 parts/week.

$$\mu_{hub} = \sum_{j=1}^{N} \mu_{FC_j} \qquad (N = \#\ of\ Supported\ FCs)$$

The variance, represented by Var(.), of two independent random variables can be combined through:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

As the utilization of parts between sites are independent of each other, part usage are not correlated and Cov(X,Y) = 0. This simplifies the above equation to a more manageable equation, which can be expanded to capture a wider range of parts.

$$Var(X+Y) = Var(X) + Var(Y)$$

$$Var(X_1, X_2, X_3 \dots X_n) = \sum_{i=1}^{n} Var(X_i) \qquad (n = Total\ number\ of\ variables)$$

The standard deviation for any given part within the hub can thus be calculated as

$$\sigma_{hub} = \sqrt{\sum_{i=1}^{N} \sigma_{FC_i}^2} \qquad (N = \#\ of\ Supported\ FCs)$$

The new lead time for a centrally stocked part is now the maximum time between the supplier to the FC (through the FC):

$$Lead\ Time_{Hub} = Lead\ Time_{Supplier\ to\ Hub} + \max(Lead\ Time_{Hub\ to\ FC})$$

Once these parameters have been calculated, it is possible to determine the (R,Q) inventory policy required within the hub. This is done in the same way as for individual FCs. The critical factor, k, is based on the highest criticality of the part seen in all FCs.

$$R = \ (\mu_{hub} * Lead\ Time_{hub} + \ \sigma_{hub} * \ \sqrt{Lead\ Time_{hub}} * k)$$

$$Q = \sqrt{\frac{2 * \mu * Fixed\ Cost\ per\ Order}{Weekly\ Holding\ Cost}}\ (or\ 1, if\ base\ stock\ model)$$

However, the (R,Q) model was developed for supply chains with large, varying demands, and simply applying the (R,Q) model will result in several issues for parts with low weekly demands. For example, imagine a part that is located at two sites, each with identical demand of 0.1 units/week (assume negligible SD) and a lead time of 4 weeks. The table below shows the calculation for the reorder point for each of the FCs and the hub.

	FC$_1$	**FC$_2$**	**Hub**
μ	0.1 parts / week	0.1 parts / week	0.2 parts / week
σ	0	0	0
Lead Time	4 weeks	4 weeks	4 weeks
R	0.4 ≈ 1 part	0.4 ≈ 1 part	0.8 ≈ 1 part

Table 6.3: Example of Inventory Policies for One Part

Recall that the inventory level of the hub consists of its own inventory, as well as all inventory downstream of it. If FC$_1$ and FC$_2$ both had one part on hand, the hub would consider its inventory level for that part to be at two (although it has no parts on hand). As such, the hub would not replenish its inventory until a part is needed downstream, but it is not able to fulfill the order (which would lead to downtime).

It is not possible to modify the (R,Q) model such that the hub's inventory is considered by itself. This would result in double counting of inventory, and typically lead to higher inventory levels than in a single-echelon system. This means that modifications have to be made to the (R,Q) model for this system.

For any given part in an FC, the lowest steady-state inventory level at any point is time is right before it hits its reorder point (i.e. R+1). As such, for any combination of warehouses, the lowest possible inventory level in the network (excluding the hub) is given by:

$$Lowest\ Inventory\ Level = \sum_{i=1}^{N} R_{FC_i} + 1 \qquad (N = \#\ of\ Supported\ FCs)$$

At the very minimum, the hub's reorder point should be at this value to ensure that it restocks prior to any of the downstream FCs requiring inventory. This is true for both (R,Q) and (S-1, S). Additionally, the hub must hold sufficient inventory to fill the expected cycle stock over the lead time, or the highest reorder quantity that could come in from a particular FC. This requires the warehouse R to be, at a minimum:

$$R_{hub} = \max\left(\left(\sum \mu_{FC}\right) * LT_{hub}\ , max(Q_{FC})\right) + \sum(R_{FC} + 1)$$

However, the (R,Q) model works in most cases. The above formula is only required to prevent stock-outs at the hub, and is only required when the (R,Q) model prescribes a restocking level that is **lesser** than this minimum quantity. In this instance, the goal is not to further lower inventory levels, but to ensure there is sufficient safety built into the model. Accordingly, the reorder point (R) for any part in the hub is given by:

$$R_{hub} = \ \max$$

Where the second term is the standard formulaic representation of the Reorder Point.

It is now possible to compare the two scenarios (multi-echelon vs single stage) for every part, to determine the optimal stocking options through the following objective function:

$$\min\left(Price * \left((R_{hub} + Q_{hub}) - \sum_{FC_1}^{All\ FCs} (R_{FC} + Q_{FC})\right), \Delta\, Shipping\ Cost\right)$$

However, this is a time-consuming process, as it requires calculating the shipping cost for both scenarios (see 6.4.2) before deciding on a stocking configuration. It is possible to simplify the objective function and use a configuration that would minimize the total number of parts across the network as shown below.

$$\min\left(\sum_{FC_1}^{All\ FCs} (R_{FC} + Q_{FC}), (R_{hub} + Q_{hub})\right)$$

Although the second objective function is less accurate, tests on Amazon data showed that the latter process stayed within $\pm 5\%$ of the total value calculated using the first objective function. Due to the price sensitivity of the data, specific outcomes cannot be shared. However, both objective functions are presented to show both options available (where the user has to decide between precision and speed). This study assumes that parts are stocked based on minimizing parts rather than total cost (i.e. the latter objective function presented above), and is able to identify where each part should be stocked.

Regardless of the method chosen, this allows the model to determine which parts should be stocked at the FC (and ordered directly from the supplier) and which parts should be stocked centrally (where FCs order from the hub). Without considering shipping costs, correct identification of where parts should be stocked could reduce Amazon's total inventory volume by approximately 14%.

The table below shows the proportion of parts that are stocked centrally vs at an FC-level for both the (R,Q) inventory policy and the (S-1, S) policy. As expected, the (S-1, S) policy pushes more parts towards centralized stocking (and thus keeps fewer total parts kept on-hand) but it is expected that this reduction in inventory would result in higher shipping costs due to more shipments between the hub and FCs (Again, data has been masked, but ratios kept relatively consistent).

	% of Parts (Centrally Stocked)	% of Parts (FC-Level)	Total Value of Parts
(R,Q)	43.71%	56.29%	$112,392,500
(S-1, S)	58.91%	41.09%	$100,000,000

Table 6.4: (R,Q) vs (S, S-1) Breakdown for Multi-Echelon System (No Shipping)

6.4.2. Shipping Costs

Once the optimal storage location for each part is known, it is possible to calculate to calculate the expected incremental shipping cost for each hub. Any other costs (e.g. labor and real estate) are independent of shipping costs and can be considered once the incremental costs are known.

Only incremental shipping costs (from hub to FC) are considered as it is assumed that the current supplier to FC costs will be fairly equivalent to the new supplier to hub costs. It is likely that nodal warehousing will lower shipping costs from supplier to hub (as suppliers only have to ship to the hub) which would provide economies of scale. However, this provides a conservative estimate that would not overstate potential savings.

In order to calculate the incremental shipping costs, the number of shipments per week of all centrally stocked parts has to be known. This is given by:

$$\# \ of \ Shipments \ per \ Week = \frac{Average \ Weekly \ Demand \ (\mu)}{Q}$$

The incremental annual shipping cost for every part that is stocked centrally can then be calculated based on the known distances between hubs and FCs and the total number of shipments per week. Assuming a base shipping price of $10 per package and 52 weeks per year, the annual shipping costs for any given part comes to:

$$\begin{aligned} &\mathit{Annual\ Shipping\ Cost} \\ &\qquad = \mathit{Base\ Shipping\ Price} * \mathit{Zone\ Multiplier} * (\#\ \mathit{of\ Shipments\ per\ Week}) \\ &\qquad * (\#\ \mathit{of\ Weeks}) \end{aligned}$$

However, in order to assess the efficacy of a multi-echelon network, this annual shipping cost must be converted into a perpetuity so that it can be compared to the immediate savings from a reduction in inventory in terms of net present value (NPV).

$$\mathit{Perpetuity} = \frac{\mathit{Annual\ Shipping\ Cost}}{\mathit{Cost\ of\ Capital}\ (\%)}$$

If a perpetuity is calculated for every part that is centrally stocked, it is possible to calculate the incremental cost of implementing a multi-echelon system through the sum of all the incremental shipping perpetuities. This is shown in the table below (data normalized such that (R,Q) perpetuity cost is $8MM)

	# of Shipments	**Perpetuity Costs**
(R,Q)	1,223	$8,000,000
(S-1, S)	2,421	$15,574,238

Table 6.5: Total Incremental Shipping Costs (Perpetuity)

As expected, the (R,Q) is significantly cheaper than the (S-1, S) model as it sends almost half as many packages. However, the formula above assumes that each package is sent separately (i.e. sending another package is an increase in price of 100%). In reality, parts going to the same site will be sent together and, as shown in the shipping pricing table above, combining shipments (i.e. adding an additional 10lbs) increases costs by ~30% rather than 100% - a significant reduction in shipping costs.

For the sake of being conservative, the model assumes 40% increase when “combining” a package. The first shipment that an FC receives every day is charged the “full” price of $10. Any subsequent part required in the same day is added on at a cost of $4 as opposed to paying the full delivery cost for a new package.

Based on the above, the shipping cost can be approximately by assuming that, for a $10 shipment, $6 (60%) is the cost of sending an empty box, and every addition of 10lbs (one shipment) is an additional $4 (40%).

In order to simplify the calculations, it is assumed that all parts are reordered as soon as they hit their reorder point, FCs will reorder all parts at or below their reorder point from the hub on a daily basis. For any given site, the weekly shipping costs is then given by:

$$Shipping\ Cost = \ Days\ in\ Week * (((0.6 * Base\ Shipping\ Price) * Zone\ Multiplier) \\ + ((0.4 * Base\ Shipping\ Price) * Zone\ Multiplier * \#\ of\ \ Daily\ Shipments))$$

The zone multiplier helps to convert the base shipping price into a shipping price based on the number of zones between the hub and the FC. Annual shipping costs are thus calculated by the formula below:

$$7 * 52 * \sum_{i=1}^{All\ FCs} (6 * Zone\ Multiplier + 4 * Zone\ Multiplier * \#\ of\ \ Daily\ Shipments)$$

Using this formula to calculate shipping costs accounts for the benefits of pooled shipping and lowers the total shipping perpetuity costs quite significantly. Note that the reduction for the (S-1, S) inventory policy is much higher than the savings for the (R,Q) inventory model.

	# of Shipments	Total Perpetuity Costs (Single Shipments)	Total Perpetuity Costs (Pooled Shipments)
(R,Q)	1,223	$8,000,000	$6,858,576
(S-1, S)	2,421	$15,574,238	$9,956,541

Table 6.6:Shipping Costs with Updated Pricing Model

The perpetuity costs from Table 6.6 can be added to the total value of parts from Table 6.4 to compare the two different inventory policies for this particular scenario (See Table 6.7).

	Shipping Costs	Total Value of Parts	Total Cost
(R,Q)	$6,858,576	$112,392,500	$119,251,076
(S-1, S)	$9,956,541	$100,000,000	$109,956,541

Table 6.7: Inventory Value and Shipping Costs

From the table above, it becomes clear that the inventory reduction from using the (S-1, S) inventory model is more than able to make up for the perpetual increase in shipping costs associated with the base stock model. As such, it makes sense for FCs to stock parts using the (S-1, S) inventory model, whilst the hub continues to manage parts using a (R, Q) inventory model (the hub does not benefit from the same benefits as FCs since they interface with multiple suppliers).

Although the model currently proposes that all parts at the FC are stocked using the (S-1, S) model, some parts may still benefit from being stocked using the (R,Q) model (e.g. parts with high demand). However, this adds a layer of complexity to the system, and should only be considered once the EAM usage data has been thoroughly verified and its accuracy can be assured.

6.4.3. Frequency of Delivery

In all instances above, it is assumed that hubs deliver parts to FCs on a daily. If shipments are delayed (e.g. delivered every two days), this would allow for more parts to be consolidated and sent together, further reducing shipping costs. However, this study finds that daily delivery results in better performance and lower total costs.

This is modelled by adding the shipping delay to the lead times of the parts (note that this may change whether parts should be stocked centrally or at an FC-level). For example, if it takes 5 days for a part to go from the hub to the FC, a shipping frequency of two days would increase that lead time to 6 days. Intuitively, if an FC requires a part on Day 1, but shipments will not happen until Day 2, it will take 6 days before the FC receives its parts.

It is acknowledged that this is an upper bound (since requests could actually fall on the day of shipment, and will not have to wait the full frequency duration). For example, some parts will be ordered on the day of shipment, and will still only take 5 days to arrive at the FC. However, to maximize conservativeness of estimates, it was chosen to apply the safety factor to all parts.

Table 6.8 and Figure 6.3 below shows how the shipping costs drop as the shipping frequency is increased (assuming base stock model).

Shipping Frequency	Shipping Costs (Perpetuity)
1 day	$9,956,541
2 day	$8,093,118
3 day	$7,471,976
4 day	$7,161,406
5 day	$6,975,063
6 day	$6,850,836
7 day	$6,762,102
14 day	$6,495,898

Table 6.8: Shipping Frequency to Shipping Costs

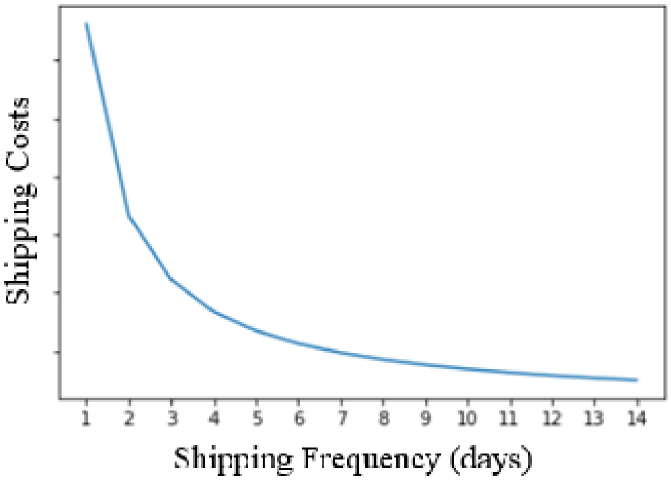


Figure 6.3: Shipping Costs vs. Shipping Frequency (days)

As expected, shipping costs decrease as the delay increases as more parts are pooled together in each shipment. The majority of savings occur if shipments are delayed between 1-4 days. This is also as expected as every additional day of delayed shipment adds proportionally fewer parts to the parts already being shipped, reducing the impact of pooled shipping.

However, there is a tradeoff in that the longer duration between shipments, the more inventory that sites will have to hold to cover that additional delay. As such, delaying shipments will decrease shipping costs, but will also increase the lead time (and total inventory required on-hand) for that part. As such, delaying shipments only make sense if the reduction in shipping costs is greater than the increase in inventory.

Due to the high service level requirements, all inventory planning for this study was based on the worst-case scenario (i.e. for a 3-4 day lead time range, the inventory model would plan for a 4 day lead time) and any shipment delays will require FCs to hold additional inventory to accommodate the extra lead time.

Table 6.9 below shows how inventory and shipping costs change when a shipping delay is added to a one-hub configuration. A one-hub configuration has the most to gain from delayed shipping (lowest inventory and highest shipping costs).

Shipping Frequency	Total Value of Parts	Shipping Costs	Total Costs
1 day	$100,000,000	$9,956,541	$109,956,541
2 days	$102,041,918	$8,093,118	$110,135,036
3 days	$103,892,373	$7,471,976	$111,364,349
4 days	$105,575,502	$7,161,406	$112,736,908

Table 6.9: Total Cost vs Shipping Frequency

From the table, it becomes clear that it is not worth postponing shipping as the additional inventory that gets moved to an FC-level (due to the increased lead time) outweighs the savings of delayed shipping. It is worth noting that this finding is independent of changes to shipping costs as the model uses the “upper bound” for shipping (i.e. third-party). Cheaper shipping costs will make it even better to minimize delays as it moves more parts to the hub (since inventory cost reductions will outweigh shipping costs by an even greater margin).

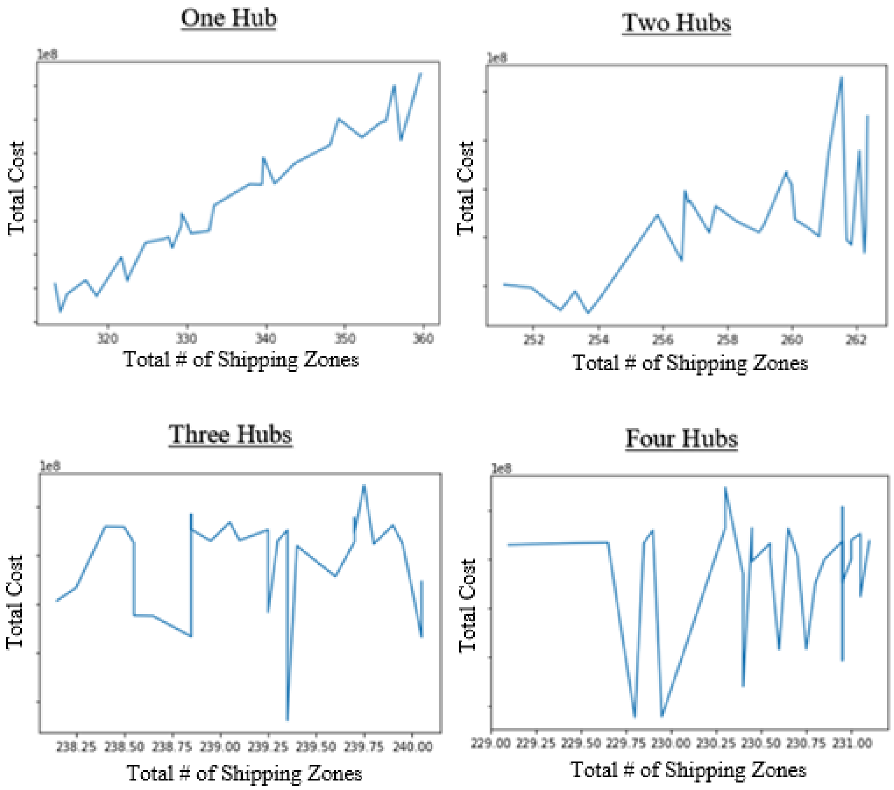
As such, in all scenarios, the optimal solution is to provide FCs with daily replenishments unless there is an operational reason to delay shipments.

6.5. Number and Location of Hubs

Another key decision that has to be made is the network’s configuration: more specifically, the number of hubs in the network and where those hubs are located. In general, as the number of hubs increase, hubs will be located closer to FCs, which will decrease lead times and shipping costs. On the other hand, increasing the number of hubs will result in less inventory being pooled at each warehouse, resulting in the network needing to carry more inventory due to greater demand variability. The selected nodal configuration should carefully balance the two factors, and aim to maximize risk-pooling within the network while minimizing lead times and shipping costs.

Figure 6.4 below show total costs (y-axis) against total shipping zones between hub and FCs (x-axis) for different hub configurations. Although shipping costs will generally increase as hubs get further away from FCs, the lead times do not change enough to lower inventory volumes. Although costs generally increase with distance, the difference was typically ±2%. Also note that minimizing zones/distance does not always minimize costs as seen by the fluctuations in total cost.

As more hubs are introduced in the system, lead times for FCs are further shortened and shipping costs decrease. However, the overall costs increase as the increased inventory levels (due to reduced pooling) outweigh the gains in reduced lead time and lower shipping costs. Total Cost



One Hub

Two Hubs

Three Hubs

Four Hubs

Figure 6.4:Cost vs Total # of Zones (Various Hub Configurations)

	Inventory Costs	Lifetime Shipping Costs	Total Costs
1 Hub	$100MM - $101.10MM	$9.96MM - $11.09MM	$109.96MM- $112.19MM
2 Hubs	$103.85MM - $104.95MM	$7.81MM - $8.04MM	$111.65MM - $112.98MM
3 Hubs	$107.14MM - $107.91MM	$6.90MM - $7.13MM	$114.04MM - $115.04MM
4 Hubs	$109.89MM - $110.88MM	$6.45MM - $6.56MM	$116.34MM - $117.44MM
5 Hubs	$111.65MM - $113.30MM	$6.00MM - $6.23MM	$117.64MM - $119.52MM

Table 6.10: Cost Ranges for 30 Potential Hub Locations

The main benefit of having more hubs comes from lower lead times to FCs. If lead times are sufficiently low (within hours), FCs could run inventory levels (for some parts) down to zero and pull parts from the hub when required. However, even a 5-hub network does not provide the necessary coverage for FCs to hold zero quantities while still meeting service level requirements.

Furthermore, since CAPEX/OPEX increase with the number of hubs, it is even more undesirable to have a large number of hubs. **<u>As such, a one-hub system was found to be optimal for the Amazon FC network.</u>** Since the differences between hub locations are quite minimal, the selection of hub location should be based on other cost factors (e.g. real estate costs etc.) rather than shipping considerations.

The table below shows a comparison between the current system, the current system with an (R,Q) inventory policy implemented, and a multi-echelon system. All costs have been normalized on a total inventory cost of $100MM for the multi-echelon system.

Hub Location	Proportion of Inventory (Pooled)	Proportion of Inventory (FC Only)	Shipping (Perpetual)	Total Costs
Current System	N/A	100%	N/A	$144,416,073
No Hub (R,Q)	N/A	100%	N/A	$131,466,159
Multi-Echelon (S-1,S)	58.91%	41.09%	$9,956,541	$109,956,541

Table 6.11: Cost Comparison of Inventory Policies

6.6. Sensitivity Analysis

Based on the model developed above, different scenarios were tested to test how certain parameters would affect the overall inventory level and shipping costs of a multi-echelon supply chain. Ideally, this would identify trends and highlight specific areas that are worth investigating further. For the purposes of this sensitivity analysis, this paper considers a single-hub located within the continental United States.

All tables in this section show results with the baseline total cost normalized to $100,000,000. Inventory levels are presented a percentage of total cost rather than an absolute value to better highlight how inventory volumes change as parameters are changed.

6.6.1. Lead Time Reduction

Lead times affect cycle stock for all parts and any reduction in lead time (whether between the supplier and the hub, or the hub and the FCs) is expected to reduce the amount of inventory held in the network.

Supplier Lead Time Reduction

The table below shows how the overall costs change as lead time from suppliers shrink. As is expected, any reductions in supplier lead time will reduce total costs as less inventory needs to be held as cycle stock.

Interestingly, a reduction in supplier lead time shifts more parts back towards the FC as opposed to central stocking. This is because parts without a high variation become less appealing to pool as lead times shrink.

Lead Time	Hub Inventory	FC Inventory	Shipping Costs	Total Costs
Baseline	57.63%	42.37%	$9,616,848	$100,000,000
½ week reduction	53.75%	46.25%	$9,319,189	$97,706,118
1 week reduction	48.97%	51.03%	$8,937,050	$95,313,955
2 week reduction	41.52%	58.48%	$8,441,276	$92,808,701
3 week reduction	37.65%	62.35%	$8,186,758	$91,442,048

Table 6.12: Sensitivity Analysis: Supplier Lead Time

Hub to FC Lead Time Reduction

Alternatively, lead times can also be reduced between the hub and the FCs. The current expectations for hub to FC shipping is shown in the table below. In this scenario, a reduction of even a day results in drops in both hub and FC inventory. Although, shipping costs increase (as the reorder point for parts is now lower) but is more than made up for by the lower inventory volumes held.

Zones	Lead Time
Zones 2-4	3 Days
Zones 5-6	4 Days
Zones 7-8	5 Days

Table 6.13: Hub to FC Lead Times

Lead Time	Hub Inventory	FC Inventory	Shipping Costs	Total Costs
Baseline [3/4/5 Days]	57.63%	42.37%	$9,616,848	$100,000,000
1 Day Reduction [2/3/4 Days]	58.61%	41.39%	$9,696,785	$98,044,223
2 Day Reduction [1/2/3 Days]	58.34%	41.66%	$9,758,269	$95,972,219
2 Day Shipping [2/2/2 Days]	58.14%	41.86%	$9,775,385	$95,223,128
1 Day Shipping [1/1/1 Day]	58.37%	41.63%	$9,812,041	$93,179,873

Table 6.14: Sensitivity Analysis - Hub to FC Lead Time

6.6.2. Cost Reductions

There are two major ways in which costs can be reduced. The first method is by lowering delivery costs between the hub and FCs (which could be accomplished through bulk shipping discounts, or internal shipping options), and the second method is through negotiation lower part prices with suppliers.

Lower Hub to FC Delivery Costs

Based on the objective function used to allocate inventory (i.e. minimize volume without considering shipping costs), lower shipping costs do not affect inventory volumes. Instead, shipping costs see an equivalent reduction in overall costs (i.e. a 10% reduction results in a 10% drop in shipping costs).

Delivery Cost	Hub Inventory	FC Inventory	Shipping Costs	Total Costs
Baseline	57.63%	42.37%	$9,616,848	$100,000,000
10% Reduction	57.63%	42.37%	$8,655,164	$99,038,315
20% Reduction	57.63%	42.37%	$7,693,479	$98,076,631

Table 6.15: Sensitivity Analysis - Delivery Costs

Price Reduction (All Parts)

The effects of an overall price reduction are as expected. Any drops in price are reflected in a similar drop in total cost. Additionally, there is a minimal shifting of inventory from hubs to FCs due to price reductions is fairly minor and does not follow any particular trend.

Price	Hub Inventory	FC Inventory	Shipping Costs	Total Costs
Baseline	57.63%	42.37%	$9,616,848	$100,000,000
95%	57.52%	42.48%	$9,486,658	$95,657,979
90%	57.57%	42.43%	$9,371,371	$91,318,452
85%	57.71%	42.29%	$9,261,992	$86,977,123
80%	57.71%	42.29%	$9,135,468	$82,614,838

Table 6.16: Sensitivity Analysis - Price Reduction

6.6.3. Criticality Service Levels

The table below shows how the total costs change as the required service level for Criticality 1 parts is lowered. In general, reductions in service level requirements will move inventory from the hub to FCs as variance has a lower impact and pooling is less effective. The observed trends are similar for Criticality 2/3 parts, and also for multi-echelon systems.

Service Level	Inventory Costs	FC Inv. Costs	Shipping Costs	Total Costs
99.9%	60.63%	39.37%	$8,905,067	$103,102,475
Baseline (99.5%)	58.93%	41.07%	$8,813,333	$100,000,000
99%	57.60%	42.40%	$8,757,773	$98,506,937
98%	56.36%	43.64%	$8,685,439	$96,949,448
95%	52.98%	47.02%	$8,567,743	$94,626,391

Table 6.17: Sensitivity Analysis - Criticality

6.7. Operations Within a Nodal System

In this system, it is imperative that orders are placed immediately once a part hits its Reorder Point and hubs fill FC's orders once a day. This is because all inventory levels are optimized based on its expected consumption vs. lead time. If a site does not replenish its parts when it hits its Reorder Point, it may not have sufficient inventory to meet future demand. Additionally, since demand profiles are combined across sites, any shortages in inventory will affect the entire system. As such, it is highly recommended that orders are managed by an automatic system that reorders all the required parts once a day.

EAM admins should not place orders for their own sites. A centrally managed system will ensure that all required parts are reordered every day. This is essential as inventory levels are based on demand profiles. If FCs try to order more parts than required, they will consume parts intended for other sites. Conversely, if FCs try to under-order, the hub may not be equipped to supply the larger compensating order at a later date (as the hub would also have under-ordered from suppliers based on the FC's previous order – this is the bullwhip effect). This could result in the hub not having sufficient inventory to meet the network's demands at a later date.

6.8. Risks and Challenges

However, there are also several risks and challenges to be cognizant of while implementing a nodal warehousing system:

Inaccurate Data: The inventory models are based on data pulled from EAM. Any inaccuracies in price, usage data or lead time will result in inaccurate stocking numbers.

Backend System: Nodal warehouses work as an entire system. As such, there has to be a reliable way to track the inventory position across the various FCs and hubs.

Lost Inventory: Currently, many RME parts get lost with other non-inventory items and may not be put into inventory for days (or end up getting stowed within the FC). Given the lower inventory levels and short turnaround times, parts have to enter the spares cage as soon as possible.

Part Counts: In a similar vein to above, EAM admins must ensure that parts are signed out every time they are used. This is to ensure that parts are reordered from the hub at the end of the day.

Inventory Turnover: In a one-hub configuration, there may be large volumes of parts to be moved on a daily basis. The hub has to be properly equipped to deal with the potential traffic volume.

6.9. Summary

Nodal warehousing presents a further opportunity to lower inventory volumes by 17% through the balancing of demand variations between different sites. As a nodal warehouse will lower the lead time of parts, the (S-1, S) inventory model performs better than the (R,Q) model at an FC level.

It was found that hub locations did not significantly alter costs. As such, hub locations should be chosen based on other factors such as CAPEX and OPEX. However, the number of hubs are significant and a one-hub network performs best for Amazon as the inventory reduction from risk-pooling outweighed the increased shipping costs.

7. Conclusion and Recommendations

The chapters above presented a way in which an inventory policy can be developed, and how a single-stage supply chain can be converted into a multi-echelon network. This chapter provides a summary of the findings, next steps for implementation, as well as potential areas for future work.

7.1. Summary of Findings

Amazon does not currently employ the usage of data in developing its inventory policies for spare parts management. Historical data on consumption of spare parts was used to develop a single-stage (R,Q) inventory model that could be applied to FCs. Through the (R,Q) model, it was found that a large number of parts (~29% of total value) are currently being under-ordered and require a higher stocking level to prevent potential stockouts. However, the inventory model was still able to reduce total inventory levels by approximately 9% by adjusting the other inventory levels to more reasonable levels.

The model then considered a multi-echelon system in order to reduce lead times to FCs, and to better position Amazon to respond to emergency stockouts. As an overall network, it was found that an introduction of nodal warehousing can reduce costs by 15%, and this saving would only continue to increase as Amazon continues to build more FCs. With a multi-echelon system in place, it was found that the (S-1, S) inventory policy performs better than the (R,Q) model at the FC-level as the network benefits from pooled shipping (the EOQ model considers each part individually, and does not consider combined shipping). Using the base stock model over the (R,Q) inventory policy is responsible for almost half of the 15% of realized cost savings from a multi-echelon network.

Once parts are stocked centrally, FCs can order any replacement parts from the central warehouse on a daily basis. It was found that shipments should not be delayed in order to consolidate deliveries, as the reduced shipping costs do not make up for the higher levels of inventory required at the FC-level.

At the network level, it was found that overall costs generally increased as the hubs increased due to the pooling effect getting diluted between more hubs. As such, a one-hub configuration performed better than other configurations, even before CAPEX/OPEX costs were considered (which would increase as the number of hubs increase). The location of the central warehouse was also found to be fairly inconsequential, with most reasonable locations performing within a $\pm 2\%$ range of each other. As such, the location of the central warehouse should be decided based on other factors such as cost and ease of implementation.

The results from this study show that implementing an inventory model, even without a multi-echelon network, would result in immediate savings. However, a central warehousing strategy would further increase those savings, while also better positioning Amazon for responding to emergencies at any FCs (due to having a central repository for any replacement parts).

7.2. Next Steps

This thesis serves as a proof of concept for the multi-echelon supply chain. The next phase of this project will involve running a pilot program to verify the data from EAM and test the validity of the model.

The next phase of this project will involve running a pilot program out of a currently empty Amazon facility in Texas. With just five FCs involved in the pilot, inventory levels could immediately fall by over $1MM. If successful, the pilot site will have the capacity to grow and support the rest of the NAFC network.

There is no specific combination of sites that should be involved in the pilot, as any form of pooling will reduce inventory levels. However, if insufficient sites are pooled, the inventory reduction may not be significant enough to outweigh the additional shipping costs. As a minimum, four medium sized sites should be used for the pilot. The FCs chosen for the pilot should be relatively large and use similar parts in order to maximize the benefits of pooling, and shipping should be done from a third-party provider (e.g. UPS) on standard ground shipping. Preliminary calculations show that a one-hub configuration supporting five FCs could deliver savings of over $1 million.

The central warehouse for the pilot program should be run out of a re-purposed Amazon site that is currently unused. This would minimize the upfront capital costs required to implement the warehouse. Additionally, an unused site has the ability to be scaled up to support more FCs over time, whereas building the hub in an operational site with excess capacity (e.g. sort center) would eventually run into capacity constraints.

Due to the high stock-out costs, the central warehouse should hold more inventory than the model recommends during the pilot. This will ensure that the system is equipped in the event that the model recommends an incorrect stocking quantity. Any usages of these extra parts should be tracked to diagnose if the incorrect recommended numbers were due to a model failure, inaccuracies in the data, or statistical chance.

In the long term, costs can be further decreased by improving data reliability to allow for less conservative assumptions, or by using the buying power of a centralized hub to negotiate bulk discounts with suppliers.

7.3. Future Work

Although the model was able to return a workable result, there are several areas in which modifications can be made to further improve the output of the model.

Parts In-Use

The model currently looks at historical data to plan future demands. However, it was unable to draw correlations between usages at different sites as there was no data on the number of parts in-use at any given time. If the number of parts in-use were known, it may be possible to identify some correlation between demand and parts in-use. This should not only improve the accuracy of the model, but opens up opportunities to use machine-learning to identify cross-site trends. Additionally, this would also allow the model to be used to predict required inventory levels at new sites (new sites don’t have demand data, so it is not possible to develop an inventory policy for them using this method at present).

Empirical Distribution

Another area for improvement is the assumed distribution used for demand curves. Although a normal distribution was used in this scenario as a compromise between accuracy and speed, it may be beneficial to compare the results to an empirical distribution which does not simplify the overall distribution.

The empirical distribution was not used in this case as there were insufficient data points to achieve the required service level with any level of accuracy, and there were sufficient doubts regarding the accuracy of the current data.

Once there is sufficient confidence in the available data, it may be worth using the daily demand data to create an empirical distribution for each part, and to use that empirical distribution to build an inventory model. This eliminates any bias that could be introduced by reducing the demand profiles to a mean and standard deviation.

8. References

Axsäter, S. (1993). *Continuous review policies for multi-level inventory systems with stochastic demand*.

Berling, P. (2008). *Holding cost determination: An activity-based cost approach—ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0925527307002605

Bimpikis, K., & Markakis, M. (n.d.). *Inventory Pooling under Heavy-Tailed Demand*.

Cachon, G., & Terwiesch, C. (2013). *Matching Supply with Demand: An Introduction to Operations Management, 3rd Edition*.

Capar, I., & Eksioglu, B. (2009). *Continuous Review Inventory Models: (Q, R) Policy*.

DeBodt, M., & Graves, S. (1983). *Continuous-review policies for a multi-echelon inventory problem with stochastic demand*.

Eppen, G. D. (1979). Note—Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem. *Management Science*, *25*(5), 498–501. https://doi.org/10.1287/mnsc.25.5.498

Federgruen, A., & Zipkin, P. (1984). Allocation policies and cost approximations for multilocation inventory systems. *Naval Research Logistics Quarterly*, *31*(1), 97–129. https://doi.org/10.1002/nav.3800310112

Goh, M. (1994). *EOQ models with general demand and holding cost functions*. https://www.sciencedirect.com/science/article/pii/0377221794901414

Halkos, G., & Kevork, I. (2012). *The classical newsvendor model under normal demand with large coefficients of variation*.

Jensen, P., & Bard, J. (2002). *Inventory Theory*. https://www.me.utexas.edu/~jensen/ORMM/supplements/units/inventory/inventory.pdf

Morse, M. P., & Kimball, G. E. (1951). *Methods of Operations Research. M.I.T. Press, Cambridge, MA. - References—Scientific Research Publishing*.

Silver, E. A., Pyke, D. F., & Peterson, R. (1998). *Inventory Management and Production Planning and Scheduling*.

Zappone, J. (2006). *Inventory Theory*. https://www.whitman.edu/Documents/Academics/Mathematics/zapponj2.pdf

Zhang, D., Xu, H., & Wu, Y. (2009). Single and multi-period optimal inventory control models with risk-averse constraints. *European Journal of Operational Research*, *199*(2), 420–434. https://doi.org/10.1016/j.ejor.2008.11.047